SOCIAL SCIENCES

US federal resource allocations are inconsistent with concentrations of energy poverty

Carlos Batlle[1,2,3], Peter Heller[1,4]*, Christopher Knittel[1,5,6], Tim Schittekatte[2,4,5]

Recent data from the US Energy Information Administration reveals that nearly one in three households in the United States report experiencing energy poverty, and this number is only expected to rise. Federal assistance programs exist, but allocations across states have been nearly static since 1984, while the distribution of energy poverty is dynamic in location and time. We implement a LASSO-based machine learning approach using sociodemographic and geographical information to estimate energy burden in each US census tract for 2015 and 2020. We then compare the allocation to states from the Low Income Home Energy Assistance Program to an optimized allocation. We allocate funds to the most burdened households, providing them with enough assistance to reduce their energy expenditures so that their household energy burden is equal to a new maximum allowable energy burden. This markedly shifts funds from the northern cold-weather states to the southern warm-weather states.

INTRODUCTION

Nearly one in three households in the United States report experiencing energy poverty (*1*). This number is only expected to rise as climate change, and US decarbonization and electrification goals lead to an increase in the cost of energy services (*2*). With growing wealth and income inequality in the United States, low-income households will bear the majority of the impacts of the energy transition despite their necessary involvement in making the transition a reality (*3*–*7*). In addition, in the face of climate change, we can expect decreased heating demand and increased cooling demand (*8*), leading to greater importance of energy poverty as a result of cooling costs. Energy poverty describes the inability of a household to adequately use sufficient amounts of electricity, heat, and other energy services due to financial constraints (*9*–*11*). Living in energy poverty has direct impacts on increased mortality, decreased physical health, decreased mental well-being, and increased isolation (*12*). Overall, nearly 10% of households in the United States kept their homes at unhealthy or unsafe (either too high or too low) temperatures. In the same year, ~20% of households reported having reduced or not purchased basic necessities to pay their energy bills (*13*). Consequently, governments can experience increased spending on social services and health care services because of energy poverty.

Here, we develop a method for estimating energy poverty levels across the United States and use these estimates to assess how allocations of federal assistance under the Low Income Home Energy Assistance Program (LIHEAP) align with the spatial distribution of energy poverty, including comparing current allocations to an optimized allocation structure where we seek to limit households' total energy burden. We build a machine learning model using an adaptive least absolute shrinkage and selection operator (LASSO) to estimate average household energy burden using sociodemographic and geographical information from the US Energy Information Administration's (EIA) Residential Energy Consumption Survey (RECS) (*13*). We select a variety of input variables based on current literature demonstrating that socioeconomic status, race/ethnicity, dwelling age, building type, education, homeownership status, and geographic characteristics correlate with energy burden across various locations (*14*–*17*). Then, using the US Census Bureau's American Community Survey (ACS) data for 2015 and 2020, we obtain the average household energy burden in every census tract across the contiguous United States. While this approach uses similar data inputs as those to fusionACS (*18*), our approach differs in function. We produce a LASSO-based machine learning approach for estimating energy burden based on household demographic and geographical information (see the Supplementary Materials).

We then use our energy poverty estimates to inform a more targeted allocation of federal assistance program funds to states. LIHEAP is the leading US federal program to help address energy poverty nationwide. LIHEAP was established in 1981 and provides states with federal block grants to assist households in paying their utility bills. This program distributes federal resources to states, and states are responsible for giving funds directly to households or utilities on the households' behalf to pay down balances. Our results reveal that current federal allocations for energy assistance, based on formulas designed nearly four decades ago, do not match well with the geographical distribution of assistance needed across the country. We find that allocating funds to states so that the energy burdens of the most burdened households are reduced to the same maximum value across the country would markedly shift funds from the northern cold-weather states to the southern warm-weatherstates.

Background and approach

Recent literature has quantified energy poverty, focusing on the United States and elsewhere (*19*–*24*). With regard to the United States, the literature focuses on specific areas or subgroups of the population (*25*–*28*), whereas our goal is to provide a comprehensive view of the entire nation and to relate this to federal assistance programs. The US Department of Energy's National Renewable Energy Laboratory (NREL) has provided the first resource, the Low-Income Energy Affordability Data (LEAD) tool (*29*), that captures a national view of energy expenditures and burden. Our machine learning–

[1]Center for Energy and Environmental Policy Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Florence School of Regulation, European Union Institute, Florence, Italy. [3]Comillas Pontifical University, Madrid, Spain. [4]MIT Energy Initiative, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]National Bureau of Economic Research, Cambridge, MA, USA.
*Corresponding author. Email: pjheller@mit.edu

based method expands upon LEAD along two important dimensions. First, the LEAD tool only presents data for one snapshot in time, which does not allow for understanding how energy burden is temporally dynamic. Second, it uses self-reported energy expenditures given only for 1 month of the year, which is not reported publicly, meaning that the estimation of annual values does not necessarily fully account for the seasonal variation in energy costs throughout the months (*29*).

To build upon LEAD, we design our model to predict household energy burden based on sociodemographic and geographic variables that can be used from surveys taken at different points in time, namely, 2015 and 2020, for this analysis. The model selects the variables that best predict energy poverty. The 2015 model selects the interaction of household heating fuels with electricity prices, household incomes, and the number of household members. Households using propane or fuel oil for heating experience the largest increase in energy burden compared to those using other fuels. In addition, decreasing household income is highly correlated with increasing energy burden. Increasing the number of members in a household also increases the energy burden compared to households with fewer members. The model for 2020 selects similar variables for increasing energy burdens. In 2020, household incomes have a larger influence over households' energy burden than in 2015. The same relation is seen for heating fuels and the number of household members. In 2015 and 2020, the household type also correlates with energy burden. Households living in multifamily units (five or more apartments, especially) have a decreased energy burden compared to those living in single-family units. A complete list of the variables selected and their relationship to energy burden can be found in fig. S3.

We then determine where concentrations of energy poverty exist across the United States by obtaining estimates of the average household energy burden in each census tract (73,057 tracts in 2015 and 84,414 in 2020). This study uses an expenditure-based metric to determine whether a household's energy burden should be classified as energy poor. Originally introduced by Boardman (*30*) in 1991, the suggested metric identifies any household that spends more than a set percentage of their annual income on energy as energy poor. The accepted threshold in the United Kingdom was 10% for many years and is still in use for Wales, Scotland, and Northern Ireland for official fuel poverty statistics (*31*). The fraction of income spent on energy services includes expenditures for electricity, natural gas, propane, fuel oil, and other vectors for household use, excluding costs for transportation. In the US context, the accepted threshold for high energy burden is set at 6%, as housing costs should not exceed 30% of income and utility costs should not exceed 20% of housing costs (*25*). Households that spend more than 10% of their income on energy are classified as experiencing severe energy burden (*32*).

RESULTS

Households experiencing an energy burden greater than 6% are classified as energy poor for this analysis. Figure 1 illustrates the distribution of the estimated average household energy burden across all census tracts across 2015 and 2020 and the change in energy burdens across the 2 years.

We observe concentrations of energy poverty in the Southeastern United States, rural Northeastern areas, communities along the southern border, and areas in the Southwestern United States,

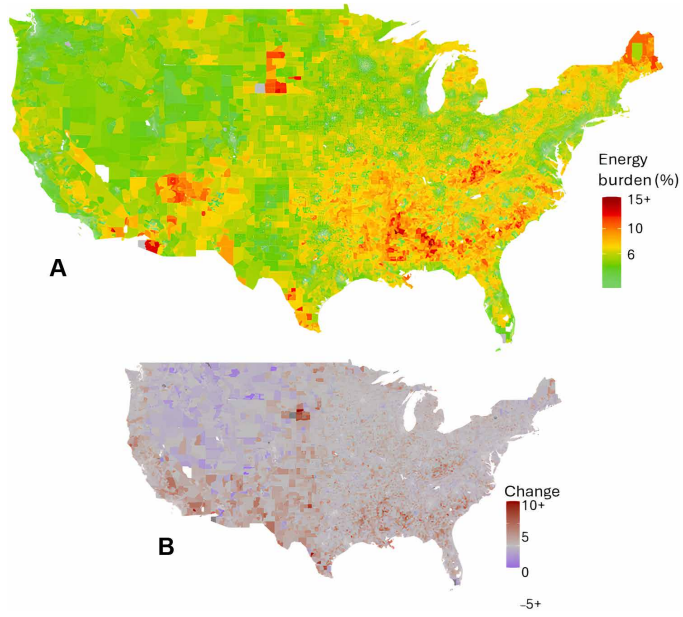

Fig. 1. Maps of average household energy burden between 2015 and 2020 and change over the period in every census tract in the contiguous United States. (**A**) Estimates of average energy burden using 2015 and 2020 US Census Bureau's ACS data in the machine learning model developed. Shades of green represent energy burdens between 0 and 6%. Shades of yellow to yellow-orange represent energy burdens between 6 and 10%. Shades of red represent energy burdens from 10 to 15% or greater. Darker shades indicate higher estimated average energy burdens. Gray areas indicate census tracts with not applicable (N/A) values. (**B**) Map representing the changes in basis points of average energy burden between 2015 and 2020 for every census tract in the contiguous United States. Blue represents tracts where the average energy burden has decreased during the period. Red represents tracts where the average energy burden has increased over time. Darker shades represent greater change. White areas indicate census tracts with N/A values.

consistent with those in (*25*). We see a small concentration of energy poverty in the south of South Dakota and in the northwest of Arizona. In 2015, Maine, Mississippi, Arkansas, Vermont, and Alabama have the highest median values of tract average energy burdens, respectively. In 2020, Mississippi, Arkansas, Alabama, West Virginia, and Maine have the highest medians. Of the census tracts in which we estimate the average household is living in energy poverty, 23% are classified as urban in 2015. This number decreases to only 14% in 2020.

We find that estimates for energy burdens increase substantially between 2015 and 2020 for areas with relatively high energy burdens, specifically in the Southeast and Southwest. In the Northwest, energy burdens decrease by up to five basis points. There are a few possible explanations for the increase in energy burdens we observe, such as increasing residential electricity rates, given the strong influence that these rates have on predicted energy burden (see fig. S3). Increasing electricity rates combined with increased adoption of air conditioning and consistent cooling degree days (CDDs) in the southern United States, as evidenced by the RECS data, create the right conditions for worsening energy burden. In addition, the COVID-19 pandemic is a likely influence that created a shock in the US economy that left many unemployed, and 46% of lower-income adults reported difficulty in paying bills (*33*).

In addition to estimating the average energy burden per census tract, our model also estimates the weighted average household energy burdens for all income brackets provided in the ACS data. This allows us to investigate the distribution of energy poverty across tracts and how this relates to median tract income, shown in Fig. 2. Lower-income tracts appear to have a greater variance in energy burden levels than tracts with higher incomes. This holds in 2015 and 2020, with larger ranges of energy burdens among all tracts in 2020. We note that both the mean and the spread of the energy burden for lower-income tracts increased substantially in 2020 compared to 2015 for tracts with incomes less than $39,999. The increased range of energy burdens for the highest median tract incomes ($60,000 to $99,999 and $100,000+) could be explained by the impact of COVID-19 on household incomes in these ranges of income.

In both 2015 and 2020, there are tracts for which the median annual household income is above $60,000, and an average energy burden is greater than 6%. The high average burden is explained by the wide variance of income levels within the tract; while the average income is above $60,000, the bottom tail is long in these tracts. Furthermore, the presence of these long left tails in income is more pronounced in 2020 compared to that in 2015. For example, in 2020, there are 4094 of the 44,891 census tracts with a median income greater than $60,000, for which our model produces an average energy burden above 6%. However, there are no households with incomes greater than $60,000 that have an energy burden greater than 6%. Instead, there is a substantial share of very low-income households in these tracts, which increases the average energy burden despite the median tract income value.

DISCUSSION

Evaluation of federal resource allocation

Next, we explore the allocation of federal resources across states to verify whether it matches the concentrations of energy poverty observed. Awareness of the need to address energy poverty in the United States arose around the early 1970's oil crisis. Since then, there have been a variety of initiatives to support consumers that can be seen as a pseudo-recognition of the problem despite no formal definition (*34*). As noted above, the leading US federal program since 1981 is LIHEAP. Figure 3 shows the allocation of LIHEAP funds across states in 2015 and 2020.

When we compare areas of high-energy burden to the allocation of federal resources, we find that funding is concentrated in the Northern United States despite the higher and increasing rates of energy poverty that we observe in the South. As climate change accelerates, we expect this misallocation to worsen. Northern areas will face lower, on average, heating burdens, while southern areas will face increasing cooling burdens. The net demand for heating and cooling combined is expected to decrease across much of the United States, except in the southeast and southern border, where net demand is expected to increase (*8*).

The misallocation can be explained by investigating the formulas used for the allocation of energy assistance. There are two formulas: an "old" and "new" formula. The old formula comes from the distribution to states based on LIHEAP's enactment in 1981, and Congress took the directly from the Low Income Energy Assistance Program, LIHEAP's predecessor. This old formula uses a variety of factors, including residential energy expenditures, heating degree days (HDDs), and household income. It did not account for cooling needs. Congress established the new formula during LIHEAP's 1984 reauthorization. The new formula made two changes. First, it uses both HDDs and CDDs and treats them symmetrically. Second, it requires the information used in the formulas to be the most up-to-date data available to the Secretary of Health and Human Services.

Despite the new formula's goal of better targeting resource allocation and treating heating and cooling needs the same, as part of the compromise to adopt the new formula, two "hold-harmless" provisions in the federal statute were also adopted: the hold-harmless

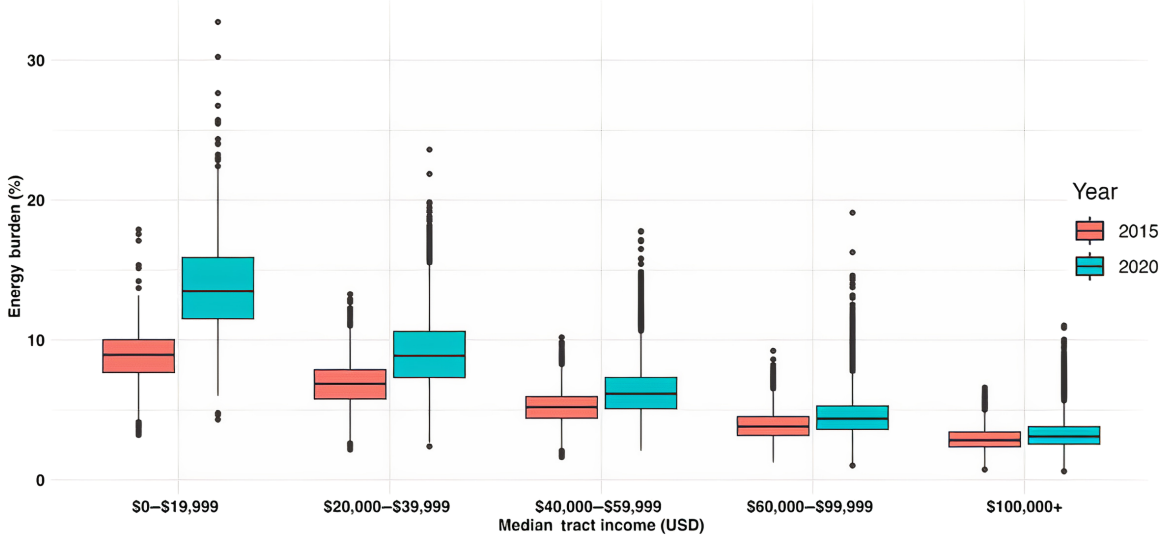

Fig. 2. Energy burden by median tract income, 2015 and 2020. Box and whisker plots for 2015 (red) and 2020 (blue) show the range of estimated average household energy burdens for census tracts with median incomes in each of the five income brackets ($0 to $19,999; $20,000 to $39,999; $40,000 to $59,999; $60,000 to $99,999; and $100,000 or greater). The five income brackets are chosen to align with available data from both 2015 and 2020. $n = 71,767$ for 2015 and $n = 82,754$ for 2020. This is a result of an increased number of census tracts in the 2020 data and accounts for removal of census tracts that return an N/A value.

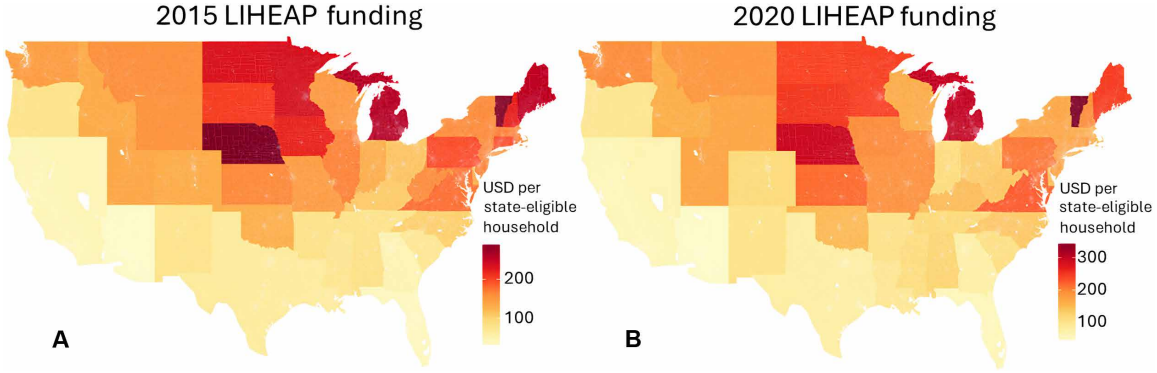

Fig. 3. Federal allocations of LIHEAP funds, 2015 and 2020. (**A**) Map of federal allocations of the LIHEAP in 2015. The total funding for each state is displayed as the US dollar amount per state-eligible household within each state. (**B**) Map of federal allocations of LIHEAP in 2020. The same units are used as in (A).

level and the hold-harmless rate. The hold-harmless level applies when LIHEAP appropriations are between the equivalent of a hypothetical fiscal year (FY) 1984 appropriation of $1.975 billion and $2.25 billion. In this scenario, any state that would receive a percentage of funds less than the old allotment percentage is guaranteed to receive, at minimum, the amount of funds that they would have received in the hypothetical FY 1984 appropriation. The hold-harmless rate is for years in which the LIHEAP appropriation exceeds $2.25 billion. In these years, states that would receive less than their old allotment percentage and less than 1% of the total appropriation must receive the percentage that they would have otherwise received if it were calculated at a total appropriation of $2.14 billion. In summary, the hold-harmless level guarantees a certain amount of funding, and the hold-harmless rate ensures a certain share of the total funding. These hold-harmless provisions create stickiness in the allocations across states, driving a wedge between the goal of the new formula, which is to treat heating and cooling the same, and actual allocations.

Clearly, determining funds available to each state is a complex process. To add to this complexity, Congress has decided several times in previous years to use language in the appropriations legislation that circumvents statute and allocates the total funds using a mixture of the old and new formula percentages (*35*). Despite the requirement that all funds be distributed according to the new formula when the appropriation is large enough, between 2009 and 2019, Congress continued to override the statute and allocate more than 80% of each year's total funding through the old formula. In doing so, federal resources continue to be allocated in a manner consistent with the way they were over three decades prior, reflecting the need to assist cold-weather states during fuel price shocks; however, conditions have changed since 1984, and the new formula that is intended to account for this change does not appear to be redistributing funds in a manner consistent with the concentrations of energy poverty that we identify. Adjusting the hold-harmless protections or allocating funds based on a new, dynamic formula that accurately accounts for the changes in conditions at yearly intervals is urgent to distribute funds to states where households require more assistance.

Optimized federal resource allocation alternative

We conclude by exploring one new allocation option for adjusting federal resource allocations to states based on our machine learning results. In the Energy Policy Act of 2005, Congress required that each state must show that the highest level of assistance is given to households with the highest energy costs and lowest income (e.g., highest energy burdens) (*36*). Following this requirement, we design a formula that takes the given budget for assistance and targets households with the highest energy burdens, as indicated by our model results, at the national level. We allocate funds to these households, providing them with enough assistance to reduce their energy expenditures so that their household energy burden is equal to a new maximum allowable energy burden.

While we use energy burden as the basis for the proposed formula in this analysis, it is important to note that there is an important avenue for future research related to the metric used to quantify energy poverty. In measuring energy poverty, there are three main methods explored in recent literature: expenditures, direct measurement, and consensual survey data (*10*, *11*, *22*, *24*–*26*, *37*–*39*). These different methodologies capture important relationships between energy costs, income, and energy usage data that provide alternative insights into how households experience energy poverty (e.g., a household that is not identified as energy poor based on their burden but maintains set heating and cooling temperatures that vary from the ideal or healthy levels). In designing a formula for resource allocation, the lack of data availability and complexity associated without having an official definition leaves space for these alternative metrics to be explored.

The maximum allowable energy burden is equal across the nation and is determined by the size of the assistance program budget. This approach ensures that no single household receives an assistance amount that would lower its energy burden beyond the maximum energy burden experienced by any household across the country (see the Supplementary Materials). Consequently, this method is akin to "shaving off" the peaks of energy burden across households. A visual representation of the old, new, and our optimized formula allocation methods can be found in the Supplementary Materials.

In doing so, we treat heating and cooling needs the same. This is consistent with the congressional goals stated in the Energy Policy Act of 2005 and the new LIHEAP formula. While these appear to be the goals of Congress, one could argue that heating needs are more important than cooling needs. Data from the Center for Disease Control and Prevention (CDC) show that, from 2006 to 2010, about two-thirds of temperature-related deaths were attributable to cold

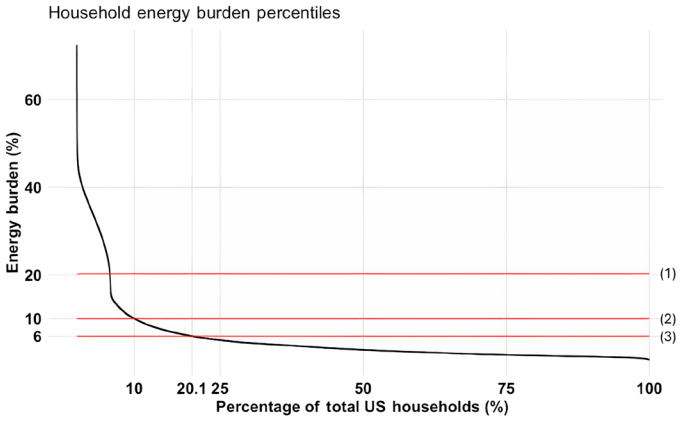

Fig. 4. Assistance funding required to achieve different levels of maximum nationwide energy burden. This plot shows the complementary cumulative distribution function illustrating the distribution of household energy burdens across the United States. The figure illustrates the percentage of US households that have an energy burden at or above a given value. For example, 10% of US households have an energy burden greater than or equal to 10%. The three red lines show the ability of different federal funding amounts to reduce the maximum energy burden experiences by any households to a given value. (**1**) This line illustrates what distribution of the current LIHEAP budget [$4.68 billion US dollars (USD)] could lower the maximum energy burden to (20.3%). In this scenario, 5.7% of US households would receive assistance. (**2**) To reduce the maximum energy burden experienced by any household to 10% (severe energy burden), $9.75 billion USD would be required and 10% of households would receive assistance. (**3**) To eradicate energy poverty in the context of this paper (no more than 6% energy burden), $17.9 billion USD would be required and 20.1% of households would receive assistance.

weather rather than heat exposure. The literature focuses on the mortality effects of extreme cold and heat (*40*–*42*). While this is useful, they do not identify to what extent mortality is prevented by specific interventions, including subsidizing heating or cooling. We come back to this issue below.

We can then determine how a given set of funds would reduce the maximum energy burden experienced across the contiguous United States, shown in Fig. 4. We see that if we were to allocate the same budget as the 2020 LIHEAP allocations, $4.7 billion US dollars (USD), then the new peak energy burden across all households would be 20.3%. In this scenario, there would still be 10% of households living with a severe energy burden; however, none of those households experience an energy burden greater than 20.3%. If the goal of energy assistance in 2020 was to make it such that no household is living in energy poverty (burden greater than 6%), then the assistance program would need $17.9 billion USD in funding, implying a nearly fourfold increase compared to the 2020 LIHEAP budget.

After funds have been allocated to households to reduce the national peak energy burden, we then calculate the total funds that each state would receive under this design. In addition, we are able to show how the distribution and severity of energy burden would change in each census tract. Figure 5 illustrates the difference in how funds are distributed and in the average energy burden in each tract based on the 2020 allocation of funds versus the new formula that we design. We see that the amount of funds per energy-poor household markedly shifts from northern states to the eastern and southeastern United States, where our results indicate that energy poverty is concentrated. There is also a more equal distribution of energy burdens under the federally optimized allocation of assistance funds, whereas, in the current system, states in the north are able to nearly eradicate energy poverty, but southern states still see broad-sweeping concentrations of severe energy burdens.

There are two general ways policy-makers could use this analysis to update allocations. The first would be to directly implement our machine learning approach, along with the optimization algorithm. However, such additional analysis might be difficult to implement. A second way, therefore, would be to use the data they already collect to calculate the "new formula" but change the weights used within the formula. To calculate the new formula allocations, HHS could use state-level data on population, percent of the population below the poverty level, HDDs, CDDs, average home energy expenditures, average home energy expenditures for heating, and average home energy expenditures for cooling and assign weights to construct allocations. These variables are similar to those that are already collected to determine allocations. To get a sense of the reduction in efficiency of the latter method, we regressed the optimal state-level allocations on this set of variables using both a purely linear model and a log-log model [note that we calculate the log-log model's coefficient of determination (R^2) after transforming the predictions and ground truth back to their linear form so that the two R^2 calculations are comparable]. The R^2 values of these regressions are 0.88 and 0.93, respectively, suggesting that Congress could achieve efficiency levels close to our full-optimized algorithm by constructing weights applied to the covariates used in our analysis.

We offer two caveats along with these calculations. First, it is important to note that, for the federally optimized allocation of funds to be implemented successfully, the system for administering funds to households would need to be changed. The current system relies on households submitting applications to their state governments or designated administrators and receiving an approval notice. Once approved, a household receives an assistance amount determined by the administrator, typically with a maximum amount set for either heating, cooling, or crisis assistance. In contrast, our allocation method would take the budgeted amount and provide assistance to any household that has an energy burden greater than the peak energy burden identified and provide enough assistance to reduce a household's energy expenditure so that their burden matches the peak burden. This would mean that any eligible household based on this new criterion would receive the funding. This method of proof of eligibility and automatic payment is styled after several currently operating systems in Europe. For example, Spain's electricity social bonus and thermal social bonus or Italy's social electrical bonus and gas social bonus use this strategy. France's energy check uses information from the previous year's tax returns to automatically qualify households and send assistance amounts.

The second caveat relates to the discussion above regarding potential differences in the welfare implications of subsidizing heating and cooling needs. Rather than take a stand on the relative welfare weights across the two needs, we ask the following question: How much more important would heating needs have to be for the current allocation across northern and southern states to be optimal? We discuss the details of this in the Supplementary Materials, but, effectively, we define burden as burden = cooling costs + β heat costs + other costs and search for the β value that would make the existing allocation across southern and northern states optimal. We find that would have to be 2.6, suggesting that

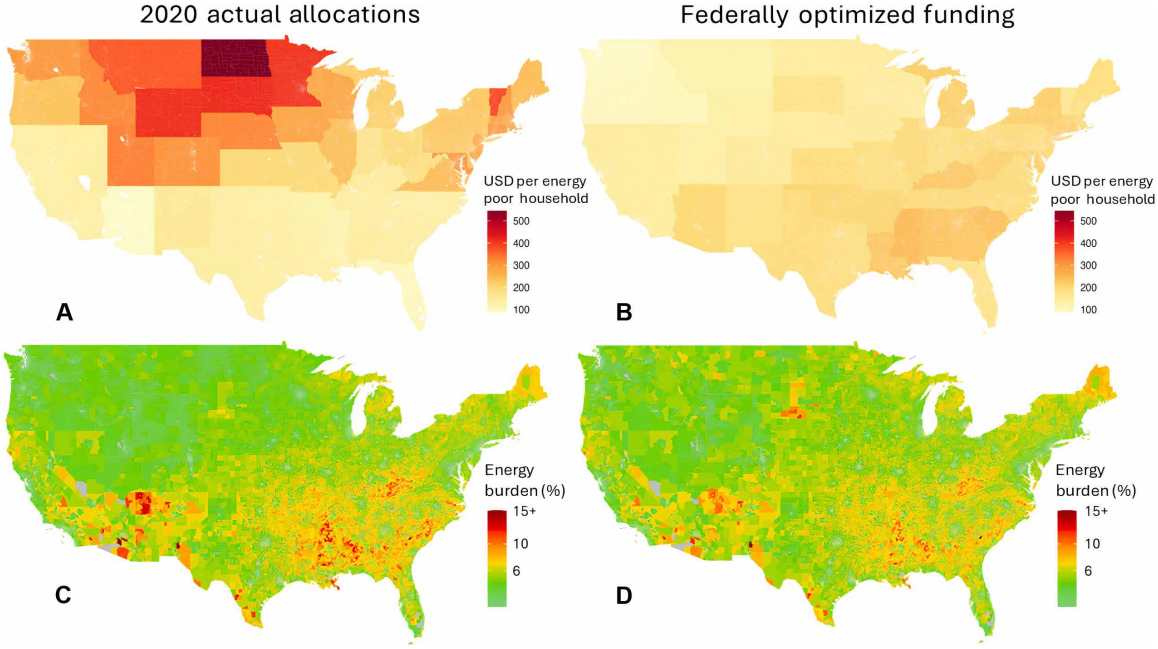

Fig. 5. Optimal allocation of federal funds and resulting energy burdens. (**A**) Map of current federal allocations ($4.7 billion USD) of LIHEAP in 2020. Units are dollars per household identified as energy poor by the 6% energy burden metric in our model. (**B**) Map of optimized federal allocations of funds to address energy burden equally across the United States, given the same budget as LIHEAP in 2020. Optimized in this scenario is defined by providing enough funds to every state in the United States such that the maximum energy burden experienced by any household is equal across the country. Units are the same as (A). (**C**) Map of average energy burden in each census tract if the 2020 LIHEAP budget were allocated to states based on current LIHEAP allocations. Households receive assistance from the state so that no household exceeds the maximum energy burden allowed given the funds received. We see that energy burdens are nearly all below 6% in the Northern United States and energy burdens remain high in the South. The maximum energy burden experienced by any household in the United States would be 29.2%. (**D**) Map of average energy burden in each census tract if dollars are allocated in a "peak-shaving" manner. In this scenario, federal dollars are allocated to each state so that the maximum energy burden that any household experiences is equivalent across the country. In this map, the maximum energy burden experienced by any household would be 20.3%.

Congress would have to value heating costs 2.6 times more valuable than cooling costs.

Energy poverty is a complex issue affecting households across the United States. It directly affects human health, social welfare, and economic success. We aim to contribute to the identification of where energy poverty exists and how it changes over time. Low incomes, household heating fuels, and the number of household members have the greatest impact on energy burdens. Between 2015 and 2020, both the mean and variance of energy burdens among all income groups increased. Our analysis finds that average household energy burdens increased from 2015 to 2020 in areas with already relatively high burdens. During this same time window, though, we observe that allocations of federal resources, through LIHEAP, remain biased to cold weather states in the North and Northeast. This is a result of antiquated formulas that are used when determining the amount of assistance given to each state. After determining a new formula for assistance allocation that would equitably reduce the peak energy burden experienced by all households, we find that funding allocations shift markedly to southern states and away from northern states. As the climate and energy markets continue to evolve, it is important to update assistance funding allocation to better match where energy burdens are highest in the United States to address energy poverty moving forward. With energy poverty concerns risking the transition to a low-carbon economy and the political inertia in increasing budgets for the assistance of energy poverty, we urge policy-makers to revise the allocation of funds to better match the distribution of assistance needed across the country.

MATERIALS AND METHODS

We use machine learning to determine how various demographic and physical characteristics are correlated with household energy burdens across the United States. Energy burden estimates allow us to identify where energy poverty may be concentrated at the census-tract level. Our analysis extends and improves upon the LEAD tool, developed by the US Department of Energy's NREL to estimate energy expenditures and burdens in several ways (*29*). The LEAD tool is designed to help local and state governments with decisions for addressing energy poverty; however, it is static in time and uses self-reported energy expenditures given only for 1 month of the year, which is not reported publicly. The reliance on 1 month implies that the estimation of annual values is not guaranteed to account for the seasonal variation in energy costs throughout the months. The sampling done by the survey must sufficiently cover all months of the year, and this is not verifiable from the publicly available data. In addition, which month is used varies across respondents. Different from LEAD, we use household-level sociodemographic and geographic data, detailed in the following subsection, from the EIA's RECS to estimate the annual energy burden. This survey is completed every 5 years, enabling us to track changes in energy burden over time.

To develop our projections at a census-tract level, we use an adaptive LASSO technique to select important variables from the RECS data to be applied to census-tract level information from the US Census Bureau's ACS. We use this methodology in six steps, illustrated by fig. S1. In what follows, we first introduce the data sources. After, we describe the machine learning approach.

Materials

We perform machine learning analysis on the RECS from the EIA (*13*). RECS reports a nationally representative sample of households to collect demographic information, physical household characteristics, and energy usage patterns. These data are joined with information from energy suppliers to produce estimates of energy services usage and costs for different end uses, including heating, cooling, and appliances. From the surveys, we select 17 variables as input variables into the model for estimating the total cost of energy services: census division, urban/rural classification, International Energy Conservation Code (IECC), CDDs, HDDs, household race, origin, highest education level achieved, age, number of household members, ownership status, annual income, household type, year built, heating fuel used, number of rooms, and number of bedrooms. We also include four additional variables, the price of electricity, natural gas, propane, and fuel oil, obtained from the US EIA's State Energy Data System (*43*). The model for 2015 also includes a variable for the length of tenure by the respondent, indicated by the year that they moved into the home.

The 2015 survey included 5686 households, representing nearly 120 million households at the time. The most recent survey, conducted in 2020 and released in 2023, has nearly three times as many survey respondents, with 18,496 households selected to represent ~123 million households across the country. Our target variable for estimation is household energy burden. There are several ways to define energy burden. For the main analysis, we define burden as the percentage of household annual income spent on household electricity, natural gas, propane, and fuel oil usage (excluding transportation).

In both the 2015 and 2020 RECS data, our initial analysis identified several outliers within each set. Specifically, some values for total household electricity consumption reported by respondents, in kilowatt-hours, reflected abnormally high consumption levels. The maximum annual electricity consumption for a household reported was 63,216 kilowatt-hours (kWh) in 2015 and 184,101 kWh in 2020. For reference, the average household in the United States consumed 10,632 kWh in 2021 (*44*). We remove outliers from the dataset each year if the reported electricity consumption is above the 99th percentile, which was 33,309 kWh in 2015 and 33,544 kWh in 2020. Households that reported having installed solar panels on their home are also removed from the dataset, as on-site generation is included in their total reported electricity consumption. Consequently, there is no way to determine how much energy was drawn from the grid versus generated on-site. Only 1.39 and 3.51% of consumers reported solar installations in 2015 and 2020, respectively. In addition, we perform a log transformation of the dependent variable, energy burden, to account for the skewness in responses.

We obtain information regarding the relevant variables for households in every census tract across the country from the ACS (*45*). Census tracts are small, statistical subdivisions of the United States containing, on average, about 4000 inhabitants. We consider ACS 5-year estimates for 2015 and 2020, which have the benefit of increased reliability for census tracts with lower populations and small population subgroups. Data for all demographic and household characteristic variables are provided at the aggregate level for the census tracts in an effort to protect anonymity. For each variable, we know the number of households that exist in each subcategory of the variable. For example, in any one census tract, we know the number of households that identify as white, Black, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, mixed race, or other. For that same census tract, we also know the number of households that live in a one-unit detached household, one-unit attached, building with two to four apartments, building with five or more apartments, or other (e.g., mobile home, recreational vehicle (RV), and boat). Critically, we do not have information on the relationship between these distributions, e.g., of the households that identify as Black, we do not know how many of them live in each of the different household types.

Methods

When using the ACS data to estimate the average household energy burden described below, we aim to obtain estimates for each of the 13 income brackets provided in both the RECS and ACS data. To achieve this without access to household-level data from the ACS, we estimate the distribution of households within input variables for which the distribution by income bracket is not given. These variables are housing type, year built, number of rooms, number of bedrooms, respondent age, number of household members, race, origin, highest education level achieved, age, and heating fuel used. To obtain these estimates, we first use the RECS data to get the actual distribution of households for each variable by income across the entire United States. We sort the RECS data by the reported household income and then sum the number of households within each variable subcategory. This allows us to calculate the proportion of households within each income bracket for each of the variable subcategories. Table S1 reports these values for the 2015 RECS data, and table S2 reports these values for the 2020 RECS.

These data are then used to simulate the distribution of households within each income bracket for a given census tract. The following example walks through the process for determining the distribution of household types by income for census tract 9352 in Butler County, Alabama, in 2020. We know the number of households in each income bracket, and we also know the number of households of each type (one-unit detached household, one-unit attached, building with two to four apartments, building with five or more apartments, or other). Using the national distribution of household types by income, we fill in the matrix presented in table S3.

The first step is to apportion the households in each income bracket into each subcategory based on the national distribution. For example, to obtain the value of $x_{1,a}$, the value of 56 households in census tract 9532 in Butler County, Alabama, that report an income between $10,000 and $14,999 is multiplied by the national proportion of households that report the same income and live in a one-unit attached type, in this case, 3.8%, obtained from the RECS data

$$x_{1,a} = 56 * 3.8\% = 2 \text{ households}$$

This is repeated for every income bracket to obtain the estimated distributions by income bracket with the census tract. The results of this step are shown in table S4.

To preserve the empirical ACS data on how many households of each type exist within a census tract, the estimated distribution of households within each type is scaled so that the sum of households in each column is equal to the total number of households of that type that exist. For example, among the estimated households that are one unit attached in each income bracket, each value is multiplied by the total number of households that live in a one-unit attached home in that tract and divided by the sum of the households allocated to that type for each income bracket

$$x_{1,a} = x_{1,a} * \frac{\text{Total one} - \text{unit attached households in tract}}{\sum_{i=1}^{13} x_{i,a}}$$

The results of this step are presented in table S5.

While this approach produces non-integer values for the number of households that exist in a subcategory, it allows us to calculate the proportion of households in each subcategory for each income bracket. Given that each subcategory is used in the regression model as a dummy variable in which the RECS data provide a zero or one value, we assume independence among each variable and use the proportion of households in the ACS data for each subcategory to estimate the average energy burden for a household in each income bracket within the tract. For example, among households in the income bracket of $10,000 to $14,999, the value in each of the five subcategories is divided by the sum of households in the income bracket

$$x_{1,a} = \frac{x_{1,a}}{\sum x_{1,a} + x_{1,b} + \ \ldots \ + x_{1,e}}$$

The results of this step are presented in table S6.

Last, for the IECC of a given census tract, guidance from the Pacific Northwest National Laboratory is used to get the climate code for each county and then applied to all census tracts within each county (*46*). We also consider CDD and HDD data from the National Oceanic and Atmospheric Administration database at the county level for 2015 and 2020 to the census tracts within the counties (*47*). We obtain estimates of electricity, natural gas, propane, and fuel oil prices from the US EIA's State Energy Data System (*43*).

To estimate the total household energy service costs across each census tract, we test several variations of regularized regression: ridge, LASSO, elastic net, and adaptive LASSO. Ridge regression, also known as L2 regularization, adds a penalty term to the linear regression cost function that discourages large coefficients, helping to mitigate multicollinearity issues. LASSO regression, on the other hand, applies L1 regularization, which encourages sparsity in the coefficient estimates, effectively selecting a subset of the most important features. Elastic net regression combines both L1 and L2 penalties, striking a balance between feature selection and multicollinearity control. Last, adaptive LASSO enhances LASSO by giving different weightage to each feature, allowing it to automatically select the most relevant predictors for the specific dataset, thereby improving model interpretability and performance. These regularized regression techniques are used to enhance the accuracy and generalization of our model in predicting household energy service costs at the census tract level.

The data provided by the RECS are used as the training and test data for each regression model. There is a combination of continuous and discrete category variables from the data. To capture potential synergistic effects between the household heating fuel variables and the price of those fuels, linear interaction variables are created by multiplying each household heating fuel variable with each of the fuel price variables. As a result, there are 92 total input variables in each 2015 model and 97 total input variables in each 2020 model. The difference in number of variables is a combination of the 2015 model including the length of tenure and the increase in discrete categories of income brackets from 2015 to 2020. In 2015, there are five income brackets (below $20,000; $20,000 to $39,999; $40,000 to $59,999; $60,000 to $99,999; and $100,000 onward). In 2020, there are 13 income brackets (below $10,000; $10,000 to $14,999; $15,000 to $19,999; $20,000 to $24,999; $25,000 to $29,999; $30,000 to $34,999; $35,000 to $39,999; $40,000 to $49,999; $50,000 to $59,999; $60,000 to $74,999; $75,000 to $99,999; $100,000 to $149,000; and $150,000 onward).

Table S7 summarizes the key metrics for each model, including mean squared error (MSE), R^2 out of sample, and the number of coefficients retained. These metrics provide valuable insights into the predictive accuracy, goodness of fit, and complexity of each model. MSE serves as an indicator of the model's predictive precision, with lower values indicating better performance. R^2 out of sample quantifies the proportion of the variance in the target variable explained by each model, offering insights into their explanatory power. In addition, the number of coefficients retained offers an understanding of the model's simplicity, as a smaller set of coefficients suggests a more interpretable model. The results in table S7 allow us to compare the regularized regression models, aiding in the selection of the most effective model for estimating household energy burden across census tracts.

Adaptive LASSO is selected as the machine learning technique for this analysis because of its ability to discover relevant predictive variables and achieve high prediction accuracy while maintaining very similar MSE and R^2 values (*48*). The process is named "adaptive" LASSO because adaptive weights are used to penalize the different coefficients in the l_1 vector. The method seeks to minimize

$$\sum_{i=1}^{n}\left(y_i - \sum_j x_{i,j} * \beta_j\right)^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j|$$

where y_i are the estimated total energy service costs, $x_{i,j}$ are the values for each RECS input variable, and $\widehat{w}_j$ are the adaptive weight vectors defined as $\widehat{w}_j = \frac{1}{\left(|\widehat{\beta}_j^{\text{ini}}|\right)^{\gamma}}$. Values for $\widehat{\beta}_j^{\text{ini}}$ are obtained through Ridge regression, which causes the technique to penalize coefficients with lower initial estimates more. This process is performed in R using the glmnet package (*49*).

We test several values of λ through a 10-fold cross-validation method. The selected value of λ is such that the error is within one SD of the minimum cross-validated error, which produces the most regularized model to improve the model generalization (*50*). Figure S2 displays the cross-validation curves alongside visualizations of the coefficients as a function of different values of λ for both the 2015 and 2020 models. From left to right, the vertical dashed lines are the log-λ values for the minimum SE and the λ for the most regularized model.

After the model has been built and tested, we use the coefficients obtained for each variable and the transformed ACS data to get estimates of the average energy burden for each income bracket in every

census tract across both 2015 and 2020. Figure S3 shows the complete set of coefficients selected in each model and the percent change in energy burden for every one unit change in their value.

To account for the log transformation of energy burden, we exponentiate the output of the model, including an extra term for the residual MSE of the fitted regression, $\frac{\hat{\sigma}^2}{2}$ in the equation below (*51*)

$$\text{Estimated energy burden} = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_n x_n+\frac{\hat{\sigma}^2}{2}}$$

where $\hat{\beta}_n$ are the estimated coefficients from the previous equation and x_n are the values for each variable using the ACS dataset.

Last, we calculate the average household energy burden for each census tract using the weighted average of the estimated household energy burden in each census tract. The weights are equal to the number of households in each income bracket taken directly from the ACS data. Once we have an estimate for the average household energy burden in each census tract, we map the contiguous United States using the geometries provided by the ACS data.

Our analysis concludes with a dynamic approach to identify the optimal allocation of assistance funds aimed at mitigating energy burdens. In this analysis, optimality is determined by the ability of the assistance funds to reach households with the highest energy burdens relative to all households in the contiguous United States. This approach is centered on evaluating the total funding necessary to lower the severest energy burdens down to an acceptable level. To achieve this, we set a national benchmark for the maximum allowable energy burden (for example, 20%) and calculate the total funding needed to ensure no household exceeds this threshold, akin to shaving off the peaks of energy burden across households. This method effectively reduces the highest energy burdens, ensuring that households facing the greatest financial strain from energy costs receive targeted assistance. If, for instance, the cap is placed at 20%, then any household with an energy burden above this figure would get enough aid to bring their burden down to the set limit. Figure S5 provides a visual representation of the old, new, and our optimized formula allocation methods.

In determining individual household energy burdens, our analysis assumes that all households in the same income bracket within a census tract have an equal energy burden. This assumption uses the simulated distribution of households within census tracts, described in this "Methods" section, to achieve the highest granularity of estimation without accessing the secure census data on individual houses.

We determine the amount of assistance that an individual household would receive by estimating their total energy service cost and reducing it so that their energy burden is reduced to the accepted threshold. The total cost of energy services is determined by multiplying a household's estimated energy burden, from the model output, by their reported income. Figure S4 shows the total program funds required to reduce the maximum energy burden across households between 50 and 6%.

In calculating the optimal allocation of funds to reduce energy burden across the country, we have chosen to not take a stance on the optimal relative weights of heating and cooling. This is consistent with the current formula used in determining allocation percentages by the Department of Health and Human Services. However, we do calculate what the relative weighting would need to be to achieve similar allocations to the North and South under the current allocations, defined as in fig. S6. Note that the distinction between North and South is decided by the states that would have received a greater percentage share under the old formula (North) and the states that would have received a greater percentage share under the new formula (South) (*35*). We first find the median burden as a result of heating expenditures, cooling expenditures, and all other energy expenditures for each state using the RECS data. We then calculate that proportion of burden that is attributable to heating, cooling, and all other energy-related expenditures for each state. For each household's burden, we determine how much is a result of heating, cooling, and other expenditures using their state-level proportions. A household's energy burden is then

$$\text{Household energy burden} = \text{cooling burden} + (\beta * \text{heating burden}) + \text{other burden}$$

We adjust the value of β to weight the heating burden more than the other burdens until the allocations to the North and the South are nearly equal to the allocations in 2020. We find that β would need to be equal to 2.6, meaning that heating burden would need to be 2.6 times more valuable than cooling costs to justify the current allocation structure as optimal. Table S8 demonstrates the different funding amounts to the North and the South as a result of different β values.

Supplementary Materials

This PDF file includes:

Figs. S1 to S6
Tables S1 to S8

REFERENCES AND NOTES

1. R. Beall, C. Hronis, "In 2020, 27% of U.S. households had difficulty meeting their energy needs" (US Energy Information Administration, 2022); www.eia.gov/todayinenergy/detail.php?id=51979.
2. P. Denholm, P. Brown, W. Cole, T. Mai, B. Sergi, M. Brown, P. Jadun, J. Ho, J. Mayernik, C. McMillan, R. Sreenath, "Examining Supply-Side Options to Achieve 100% Clean Electricity by 2035" (NREL/TP-6A40-81644, 1885591, MainId:82417, National Renewable Energy Laboratory, 2022); https://doi.org/10.2172/1885591.
3. A. Datta, "Climate change and energy vulnerability of the American poor" (Univ. of Houston, 2023); https://nlga.us/wp-content/uploads/uh-energy-white-paper-energy-burden-june-6-edit.pdf.
4. E. Johnson, R. Beppler, C. Blackburn, B. Staver, M. Brown, D. Matisoff, Peak shifting and cross-class subsidization: The impacts of solar PV on changes in electricity costs. *Energy Policy* **106**, 436–444 (2017).
5. C. G. Monyei, B. K. Sovacool, M. A. Brown, K. E. H. Jenkins, S. Viriri, Y. Li, Justice, poverty, and electricity decarbonization. *Electr. J.* **32**, 47–51 (2019).
6. T. Blanchet, L. Chancel, A. Gethin, Why is Europe more equal than the United States? HAL open science (2020); https://shs.hal.science/halshs-03022133/document.
7. S. Carley, D. M. Konisky, The justice and equity implications of the clean energy transition. *Nat. Energy* **5**, 569–577 (2020).
8. Y. Amonkar, J. Doss-Gollin, D. J. Farnham, V. Modi, U. Lall, Differential effects of climate change on average and peak demand for heating and cooling across the contiguous USA. *Commun. Earth Environ.* **4**, 402 (2023).
9. R. Barrella, J. I. Linares, J. C. Romero, E. Arenas, E. Centeno, Does cash money solve energy poverty? Assessing the impact of household heating allowances in Spain. *Energy Res. Soc. Sci.* **80**, 102216 (2021).
10. H. Thomson, S. Bouzarovski, C. Snell, Rethinking the measurement of energy poverty in Europe: A critical analysis of indicators and data. *Indoor Built Environ.* **26**, 879–901 (2017).
11. D. Charlier, B. Legendre, Fuel poverty in industrialized countries: Definition, measures and policy implications a review. *Energy* **236**, 121557 (2021).
12. European Parliament, *Energy Poverty: Handbook* (Publications Office of the European Union, 2016); https://data.europa.eu/doi/10.2861/94270.
13. US Energy Information Administration (EIA), Residential Energy Consumption Survey (RECS) - Energy Information Administration; www.eia.gov/consumption/residential/.
14. D. J. Bednar, T. G. Reames, G. A. Keoleian, The intersection of energy and justice: Modeling the spatial, racial/ethnic and socioeconomic patterns of urban residential heating consumption and efficiency in Detroit, Michigan. *Energy Build.* **143**, 25–34 (2017).

15. L. Ross, A. Drehobl, B. Stickles, "The high cost of energy in rural America: Household energy burdens and opportunities for energy efficiency" | (ACEEE, 2018); www.aceee.org/research-report/u1806.
16. J. Lin, Affordability and access in focus: Metrics and tools of relative energy vulnerability. *Electr. J.* **31**, 23–32 (2018).
17. I. Mayer, E. Nimal, P. Nogue, M. Sevenet, The two faces of energy poverty: A case study of households' energy burden in the residential and mobility sectors at the city level. *Transp. Res. Procedia* **4**, 228–240 (2014).
18. K. Ummel, M. Poblete-Cazenave, K. Akkiraju, N. Graetz, H. Ashman, C. Kingdon, S. Herrera Tenorio, A. S. Singhal, D. A. Cohen, N. D. Rao, Multidimensional well-being of US households at a fine spatial scale using fused household surveys. *Sci. Data* **11**, 142 (2024).
19. P. Mulder, F. Dalla Longa, K. Straver, Energy poverty in the Netherlands at the national and local level: A multi-dimensional spatial analysis. *Energy Res. Soc. Sci.* **96**, 102892 (2023).
20. K. Rademaekers, J. Yearwood, A. Ferreira, S. Pye, I. Hamilton, P. Agnolucci, D. Grover, J. Karásek, N. Anisimova, "Selecting indicators to measure energy poverty" (Tech. Rep. ENER/B3/507-2015, Trinomics, 2016).
21. M. Riva, S. Kingunza Makasi, P. Dufresne, K. O'Sullivan, M. Toth, Energy poverty in Canada: Prevalence, social and spatial distribution, and implications for research and policy. *Energy Res. Soc. Sci.* **81**, 102237 (2021).
22. S. Meyer, H. Laurence, D. Bart, L. Middlemiss, K. Maréchal, Capturing the multifaceted nature of energy poverty: Lessons from Belgium. *Energy Res. Soc. Sci.* **40**, 273–283 (2018).
23. B. Legendre, O. Ricci, Measuring fuel poverty in France: Which households are the most fuel vulnerable? *Energy Econ.* **49**, 620–628 (2015).
24. L. Papada, D. Kaliampakos, Measuring energy poverty in Greece. *Energy Policy* **94**, 157–165 (2016).
25. E. Scheier, N. Kittner, A measurement strategy to address disparities across household energy burdens. *Nat. Commun.* **13**, 288 (2022).
26. S. Cong, D. Nock, Y. L. Qiu, B. Xing, Unveiling hidden energy poverty using the energy equity gap. *Nat. Commun.* **13**, 2456 (2022).
27. Q. Wang, M.-P. Kwan, J. Fan, J. Lin, Racial disparities in energy poverty in the United States. *Renew. Sustain. Energy Rev.* **137**, 110620 (2021).
28. E. Dogan, M. Madaleno, R. Inglesi-Lotz, D. Taskin, Race and energy poverty: Evidence from African-American households. *Energy Econ.* **108**, 105908 (2022).
29. O. Ma, K. Laymon, M. Day, R. Oliveira, J. Weers, A. Vimont, "Low-Income Energy Affordability Data (LEAD) tool methodology" (Tech. Rep. NREL/TP-6A20-74249, National Renewable Energy Laboratory, 2019); www.nrel.gov/docs/fy19osti/74249.pdf.
30. B. Boardman, *Fuel Poverty: From Cold Homes to Affordable Warmth* (Belhaven Press, 1991).
31. S. Oxley, *Fuel Poverty Methodology Handbook (Low Income Low Energy Efficiency)* (UK Department for Energy Security and Net Zero, 2024); https://assets.publishing.service.gov.uk/media/65ccf6341d9395000c9466a7/fuel-poverty-methodology-handbook-2024.pdf.
32. Applied Public Policy Research Institute for Study and Evaluation (APPRISE), "LIHEAP energy burden evaluation study" (PSC Order No. 03Y00471301D, APPRISE, 2005); www.acf.hhs.gov/sites/default/files/documents/ocs/comm_liheap_energyburdenstudy_apprise.pdf.
33. K. Parker, R. Minkin, J. Bennett, "Economic fallout from COVID-19 continues to hit lower-income Americans the hardest" (Pew Research Center, 2020); www.pewresearch.org/social-trends/2020/09/24/economic-fallout-from-covid-19-continues-to-hit-lower-income-americans-the-hardest/.
34. D. J. Bednar, T. G. Reames, Recognition of and response to energy poverty in the United States. *Nat. Energy* **5**, 432–439 (2020).
35. L. Perl, "The LIHEAP formula" (CRS Report RL33275, Congressional Research Service, 2019); https://crsreports.congress.gov.
36. Energy Policy Act of 2005 (2005); www.govinfo.gov/content/pkg/PLAW-109publ58/pdf/PLAW-109publ58.pdf.
37. J. C. Romero, P. Linares, X. López, The policy implications of energy poverty indicators. *Energy Policy* **115**, 98–108 (2018).
38. C. Lowans, D. F. Del Rio, B. K. Sovacool, D. Rooney, A. M. Foley, What is the state of the art in energy and transport poverty metrics? A critical and comprehensive review. *Energy Econ.* **101**, 105360 (2021).
39. D. Deller, G. Turner, C. Waddams Price, Energy poverty indicators: Inconsistencies, implications and where next? *Energy Econ.* **103**, 105551 (2021).
40. N. Seltenrich, Between extremes: Health effects of heat and cold. *Environ. Health Perspect.* **123**, A275–A279 (2015).
41. A. Gasparrini, Y. Guo, M. Hashizume, E. Lavigne, A. Zanobetti, J. Schwartz, A. Tobias, S. Tong, J. Rocklöv, B. Forsberg, M. Leone, M. De Sario, M. L. Bell, Y.-L. L. Guo, C. Wu, H. Kan, S.-M. Yi, M. de Sousa Zanotti Stagliorio Coelho, P. H. N. Saldiva, Y. Honda, H. Kim, B. Armstrong, Mortality risk attributable to high and low ambient temperature: A multicountry observational study. *Lancet* **386**, 369–375 (2015).
42. J. Berko, D. D. Ingram, S. Saha, J. D. Parker, Deaths attributed to heat, cold, and other weather events in the United States, 2006-2010. *Natl. Health Stat. Rep.* **76**, 1–15 (2014).
43. US Energy Information Administration (EIA), "Independent statistics and analysis (EIA, 2023); www.eia.gov/state/seds/seds-data-complete.php?sid=US#CompleteDataFile.
44. US Energy Information Administration (EIA), "How much electricity does an American home use?" (EIA, 2022); www.eia.gov/tools/faqs/faq.php.
45. US Census Bureau, American Community Survey (ACS), Census.gov; www.census.gov/programs-surveys/acs.
46. M. C. Baechler, T. L. Gilbride, P. C. Cole, Marye G. Hefty, Kathi Ruiz, "Guide to determining climate regions by county" (PNNL-17211 Rev. 3, Pacific Northwest National Laboratory, 2015).
47. County Mapping, Climate at a Glance, National Centers for Environmental Information (NCEI); www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping/110/cdd/201512/12/value.
48. H. Zou, The adaptive LASSO and its oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429 (2006).
49. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
50. T. Hastie, J. Qian, K. Tay, "An introduction to 'glmnet'" (2023); https://glmnet.stanford.edu/articles/glmnet.html#.
51. J. Yang, "Interpreting regression coefficients for log-transformed variables" (Cornell Statistical Consulting Unit, 2020); https://cscu.cornell.edu/wp-content/uploads/logv.pdf.

Acknowledgments: We would like to thank S. Saraf for assistance in editing the R scripts during the research design process. **Funding:** This work was supported by the MIT Future Energy Systems Center. **Author contributions:** Research design: C.B., P.H., C.K., and T.S. Data cleaning, analysis, and evaluation: P.H. Conceptualization: P.H., C.K., C.B., and T.S. Methodology: P.H., C.K., C.B., and T.S. Investigation: P.H. Visualization: P.H. Funding acquisition: C.B. and T.S. Project administration: C.K. Supervision: C.K., C.B., and T.S. Writing—original draft: P.H. and C.K. Writing—review and editing: P.H., C.K., C.B., and T.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 12 April 2024
Accepted 4 September 2024
Published 9 October 2024
10.1126/sciadv.adp8183