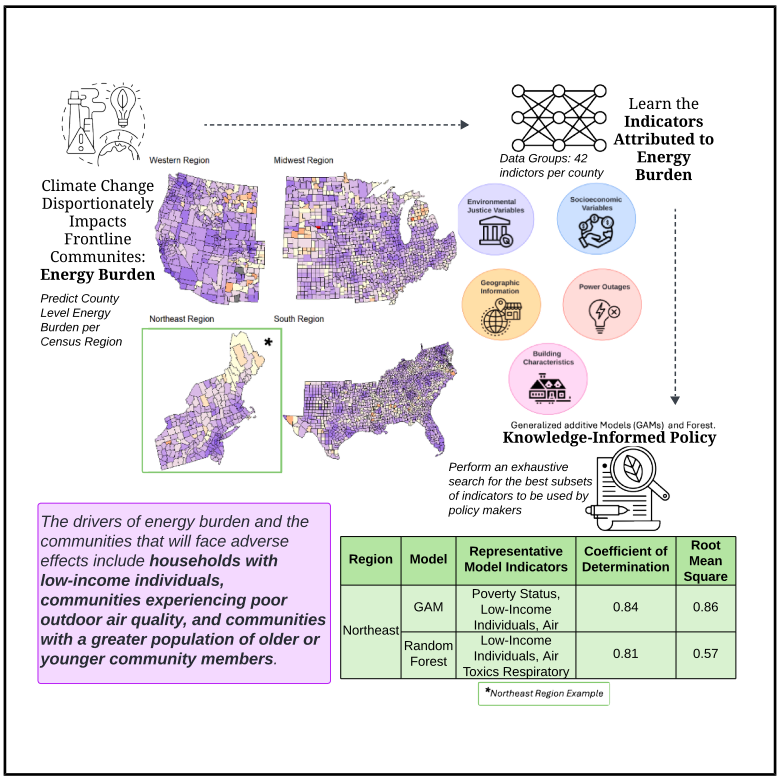
iScience

Article

Exploring the importance of environmental justice variables for predicting energy burden in the contiguous United States

Graphical abstract



Region	Model	Representative Model Indicators	Coefficient of Determination	Root Mean Square
Northeast	GAM	Poverty Status, Low-Income Individuals, Air	0.84	0.86
	Random Forest	Low-Income Individuals, Air Toxics Respiratory	0.81	0.57

Highlights

- Novel indicators
- Dataset reduction for policy action
- Energy burden prediction

Authors

Jasmine Garland, Kyri Baker, Balaji Rajagopalan, Ben Livneh

Correspondence

jasmine.garland@colorado.edu

In brief

Environmental science; Environmental policy; Energy policy

Garland et al., 2025, iScience 28, 112559
June 20, 2025
https://doi.org/10.1016/j.isci.2025.112559

CellPress

iScience

CellPress
OPEN ACCESS

Article

Exploring the importance of environmental justice variables for predicting energy burden in the contiguous United States

Jasmine Garland,[1,5,*] Kyri Baker,[1,2] Balaji Rajagopalan,[1,3] and Ben Livneh[1,3,4]
[1]The Department of Civil, Environmental, and Architectural Engineering at the University of Colorado Boulder, Boulder, CO, USA
[2]Renewable and Sustainable Energy Institute (RASEI), University of Colorado Boulder, Boulder, CO, USA
[3]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, USA
[4]Western Water Assessment, University of Colorado Boulder, Boulder, CO, USA
[5]Lead contact
*Correspondence: jasmine.garland@colorado.edu
https://doi.org/10.1016/j.isci.2025.112559

SUMMARY

The United States is one of the largest energy consumers per capita, requiring households to have adequate energy expenditures to keep up with modern demand regardless of financial cost. This paper investigates energy burden, defined as the ratio of household energy expenditures to household income. There is a lack of research on creating equitable policies for energy-burdened communities, including environmental justice indicators and community characteristics that could be used to predict and understand energy burden, along with socioeconomic status, building characteristics, and power outages, beneficial to policymakers, engineers, and advocates. Here, generalized additive models and random forests are explored for energy burden prediction using the original dataset and principal components, followed by a leave-one-column-out (LOCO) analysis to investigate indicator influence, with 25 identical indicators out of 42 appearing in the top 100 models. The generalized additive models generally outperform the random forests, with the best-performing model yielding a coefficient of determination of 0.92.

INTRODUCTION

Energy and automation are essential to the advancement of human endeavors. It is estimated that the United States (U.S.) consumes approximately 101 quadrillion British Thermal Units of primary energy in support of human development and existence; this is approximately 17% of world primary energy consumption, while the U.S. only accounts for approximately 4% of the word population.[1] Energy use impacts almost every dimension of modern society. Thus, when access to energy is limited, these impacts are compounded through housing, mobility, health, work, education, and other facets of life.[2] Governing bodies at the local, state, and federal levels have recognized the challenges climate change brings for traditionally marginalized communities. However, how to create equitable and just public policies at the confluence of climate change, energy, and disadvantaged communities remains a complex question[3] with a dearth of research.[4] Access to energy resources plays a vital role as compounded climate and electric infrastructure events occur. For example, the 2021 winter blackout in Texas left approximately 10 million people without electricity for as long as multiple days. Many news outlets reported that minority neighborhoods were disproportionately impacted.[5] While this is a multi-faceted issue, researchers at the University of California Berkeley have found that inequalities have been built into the California power grid by design. Minority groups often live in disadvantaged census blocks, which do not have equal access to distributed energy resources (DERs) such as rooftop solar and battery storage.[6]

With an increase in energy dependence and technological development, individuals are able to arbitrage energy assets with residential rooftop solar, home energy storage systems, and electric vehicle adoption. These technologies contribute to the transition to a clean energy grid, drastically changing the energy economy. However, the clean energy transition has the potential to increase the growing wealth disparity in the U.S.,[7] as many programs for solar installation or electric vehicle procurement, which low-income households often do not participate in, are financed by raising the price of electricity to all customers. Thus, asset disparities are exacerbated for low-income households as energy prices rise without an increase in energy resources or assets for their household.[6] This highlights the need for clean energy solutions and calls attention to policymakers and energy programs to ensure a just and equitable transition.

Energy justice is the confluence of energy systems and social justice, a human-centered approach to fairly distributing benefits and burdens of the current and future power and energy systems. In this paper, equity and justice will be discussed interchangeably. Within this context, justice is viewed as long-term equity. Numerous types of equity exist; however, the types of

equity that are most related to this work are defined herein. Procedural equity relates to the process of allocating resources fairly, with transparency and inclusion throughout the decision-making process. Distributive equity is the action of allocating rights and resources with fairness, which includes identifying where and when injustices occur. Thus, procedural equity is the fairness of a process, but distributional equity relates to the actual allocation of resources themselves. Intergenerational equity considers obligations to future generations, including a dimension of time and future planning. Recognitional justice is understanding different vulnerabilities and needs related to energy services and how they differ among socioeconomic groups or communities. Frontline communities are the most vulnerable to climate change and are adversely impacted by inequitable actions due to systemic and historical disparities.[8] The authors in Carley and Konisky[9] highlight the importance of procedural, distributive, and intergenerational equity for frontline communities. Additionally, McCauley et al.[10] identify procedural, distributive, and recognition justice as the three tenets of energy justice from energy production to energy consumption through policymakers' viewpoint. However, from a climate change perspective, intergenerational equity is often at the forefront.[11]

There are many constructs regarding inequalities in the energy ecosystem relating to energy poverty these are reviewed and compared in Brown et al.[2] and Tarekegne et al.[12] Energy poverty may be defined as the lack of access to basic, life-sustaining energy due to a lack of resources.[12,13] This paper will focus on energy burden; the definition of this concept is provided in Equation 1. Energy burden emphasizes the financial component of energy poverty and is considered a primary and absolute metric of energy poverty. Cong et al.[13] define a primary energy poverty metric as one that directly utilizes consumer-level information and defines an absolute metric as one that has a specific threshold for energy poverty. Given the primary and absolute characteristics of energy burden, this metric offers a standardized starting point to further understand indicators that are associated with energy poverty through energy burden versus opposing, relative, or secondary energy poverty metrics that may use weighted scoring and lack strict thresholds.[13,14]

$$\text{EnergyBurden}(\%) = \frac{\text{EnergyBills}(\$)}{\text{Income}(\$)} \qquad \text{(Equation 1)}$$

Energy burden is frequently used by the U.S. Department of Energy (DOE) and considers energy expenditures (consumption and price), household income, and affordability.[12] In this context, energy bills consider electricity, gas, and alternative fuels such as fuel oil and wood; the income is gross income. The U.S. DOE states that households experiencing an energy burden of 6% or greater are considered to have a high energy burden, and households with an energy burden of 10% or higher have a severe energy burden. These thresholds were created with the notion that a household should not spend more than 30% of the income on housing expenses, and utility costs should not exceed 20%. Utility costs do not include transportation energy or water use[15] and the cost of living for separate regions are not considered.[16]

A high energy burden can result in shutoffs and "bundled burdens" such that economic trade-offs occur, creating a cumulative risk to the household. Trade-offs include living in comprised homes and the "heat or eat" phenomena, resulting in the co-occurrence of food and energy insecurity.[17] However, co-occurrences are not limited to energy and food but include medical care, proper shelter, and other life necessities. Solutions to energy injustices have been sparse. The residential energy consumption survey found that one in three U.S. households have faced challenges paying their energy bills.[18] Although, energy burden has gained popularity in the past decade due to policies and programs such as weatherization and low-income home energy assistance programs. The respective programs are the nation's most extensive energy programs concerning low-income household energy assistance.[2] As a result, Brown et al.[19] conducted an expansive bibliometric analysis regarding energy equity in the U.S. and found 183 peer-reviewed papers and government reports published between 2010 and 2019.

A case study in Arizona showed that low-income households wait 2.6°C and 4.2°C longer in the summer months to turn on their air conditioning (A/C) compared to more financially secure households.[20] Thus, energy burdened households may put themselves at risk for adverse health effects, such as respiratory issues, exposure to indoor air pollution, lead exposure, mold growth, general thermal discomfort, and other health related issues.[21,22] In Chen et al.,[23] spatial analysis is considered to look at select counties with a high energy burden and the connection between healthcare and COVID-19, finding that high energy burden communities are significantly less likely to have health insurance. The study also links geographic features, showing that communities near each other often have similar characteristics. Further, low-income and middle-income households were more likely to experience energy poverty due to limiting energy use behavior due to increased energy bills from stay-at-home orders during COVID.[24] The authors in Buylova[25] and Bednar et al.[26] evaluate statistical modeling techniques to investigate the intersection of residential building energy use intensity (EUI), racial/ethnic households, and socioeconomic patterns in the state of Oregon and the city of Detroit, Michigan. A relationship between high residential EUIs, racial minority households, and education was detected. Further, Moore and Webb[27] models energy burden for Cincinnati, Ohio, using economic, social, and energy related metrics. Findings conclude that spatial models outperform their non-spatial counterparts and that socioeconomic variables, particularly income-related metrics, are the strongest indicators to predict energy burden.

A more researched topic relating to energy burden is environmental justice. Environmental injustice, also referred to as environmental inequality, occurs when a social group faces disproportionate negative impacts from environmental hazards, most often in the form of environmental racism. Environmental racism occurs when individuals, groups, or communities of a race, ethnicity, or color are impacted or disadvantaged by a policy or practice, intended or unintended.[12,28] Studies of environmental justice date back to the 1970s as evidence of toxic hazards impacting communities of color were brought to the forefront by social activists.[29] Since this pioneering work, the link between poor outdoor air quality and poverty has been well documented,[30] while the authors in Hauptman et al.[31] have linked low-income communities and household

lead exposure (homes built before 1960). Further, Hilmers et al.[32] studies disparities in food deserts and transportation in low-income neighborhoods. Thus, environmental justice factors are often presented as ''bundled burdens''.

Energy justice and environmental justice are constitutionally interconnected as energy justice builds on environmental justice principles from extractive economies among the energy, waste, and environmental sectors that have disproportionate impacts on low-income communities.[33] One example is the negative impacts on air quality and health from fossil fuel power plants that disproportionately impact frontline communities. While fossil fuel power plants are not the sole cause of the link between low-income and poor air quality, frontline communities would benefit significantly from a just energy transition. Additionally, environmental justice has been more thoroughly studied than energy justice or energy burden. Due to the greater maturity in our understanding, policymakers and engineers can use environmental justice indicators as a beacon of knowledge to determine areas that may experience ''bundled burdens''. Thus, for a just energy transition, shifting from an extractive economy toward a regenerative economy would allow for equitable asset allocation and community control of the local economy. Equitability in economic development is essential as treating every community equally when disparities exist harms those on the frontline.

Another area that frontline communities disproportionately feel the impacts of climate change is power outages.[34] The Environmental Protection Agency (EPA) indicates that the average power outage duration between 2013 and 2021 doubled, from 3.5 h to 7 h, and the frequency increased from 1.20 to 1.42 events per customer per year. However, power outages do not impact communities equally, as socioeconomic status is often correlated with power outage occurrence and duration.[35] This is particularly true with the increase in weather-related power outages due to unprecedented changes in climate. For instance, during the 2021 Texas freeze, census blocks with a high minority population were four times more likely to experience a power outage, as 10–11% of prominently white neighborhoods experienced a power outage, whereas 47% of minority populations experienced an outage,[36] further zip codes with a higher minority population experienced more frequent power outages.[37] Multiple studies[34,36–39] have found disparities in how minority neighborhoods experience power outages, yet the cause has not been identified.

The authors in Do et al.[34] found that socially vulnerable communities, as defined by the Centers for Disease Control and Prevention, on average, experienced power outages that were 3 times longer. Further, a one-decile drop in social vulnerability would result in a 6.1% longer power outage. Similar results are found in Garland et al.,[40] given that the rural county in the study experienced more extended power outages and predictions differed from its more heavily populated counterparts. Rural communities that experience power outages for greater durations could be a characteristic of the ''last mile'', in which power recovery often takes several days.

The exact cause of this disparity is not well documented, but considerations include aging infrastructure, geographic location, and bias in prioritizing communities with higher incomes and/or predominantly white communities. Although there are cases where resources are equal among regions, yet some communities remain more vulnerable to social and economic loss; this could be partially due to resources such as DERs and community facilities. One crucial piece of resilience in low-income communities is the willingness to pay, which often increases as income increases. Making everyone in a region pay a flat rate would add additional burdens to those already experiencing energy burden or financial difficulties.[41] Thus, understanding the characteristics of communities experiencing more frequent and extended outages is vital to implementing solutions that provide equal access to reliable energy without disproportionate burdens.

Machine learning (ML) has the potential to advance public policy when implemented through a human-centered lens. As in Coyle and Weller,[42] decision making for public policy with the use of ML is reviewed in terms of learning relationships between data inputs (features or, in this case, energy burden indicators) and decisions (outputs). Relevant to this study are the post hoc interpretations of the decisions. Post hoc interpretations are completed after the study or model has been constructed and results have been produced, which includes the interpretation of feature importance. Further, meaningful indicators, including social indicators, are vital to knowledge-informed policies. Knowledge-informed policies are knowledge influenced, meaning indicators have been thoroughly understood before creating the policy. Otherwise, it is purely a political policy.[43]

Understanding the importance of indicators in ML models is essential, as the numerous variables of interest to policymakers typically have complex interactions. For instance, Bell et al.[44] found that zip code was the most important feature in predicting housing prices for tax purposes, which was not particularly helpful to their domain. However, they also found that zip codes were strongly correlated with race. The use of an indicator such as race could breach the Fair Housing Act of 1968. Thus, the zip code was removed from the models. This highlights the importance of model interpretability, explainability, and human interference in ML for public policy.[45] As such, modeling methods that aid in a deeper understanding of the multi-dimensional nature of energy and environmental justice and the power system were prioritized in this study.

Previous studies have explored machine learning for energy justice; for instance, Ghorbany et al.[46] and Ghorbany et al.[47] consider passive building design strategies and energy burden using multiple machine learning methods. Ghorbany et al.[47] include social demographic variables among building characteristics and found machine learning methods to offer the best results and that passive design contributed to model performance, while Ghorbany et al.[46] found that passive design reduces energy burden and that location was a key indicator. In Spandagos et al.,[48] energy burden is studied across the European Union, and it is found that the random forest model was one of the best-performing models. Random forest models are gaining popularity, having been considered in multiple studies relating to climate,[49,50] sustainability,[51] and energy poverty.[48,52–55] However, these studies do not focus on the entire U.S. and do not consider the combination of data groups and variables considered in this study, which are found in Table 1, particularly the environmental justice and power outage variables defined within the context of this paper. While other work has explored

Table 1. Data groups

Data group	Data description
Community resilience for equity and disasters[56]	Estimated number of individuals with zero risk factors, estimated number of individuals with one-two risk factors, estimated number of individuals with three plus risk factors.
Socioeconomic variables[57]	Total population count, population for whom poverty status is determined, count of households in linguistic isolation, count of people of color individuals, count of low-income individuals, count of individuals age 25 or over with less than high school degree, count of individuals under age 5, count of individuals over age 64.
Environmental justice variables[57]	Count of housing units built before 1960, diesel particulate matter level in air, air toxics cancer risk, air toxics respiratory hazard index, traffic proximity and volume, indicator for major direct dischargers to water, proximity to national priorities list (NPL) sites, proximity to risk management plan (RMP) facilities, proximity to treatment storage and disposal (TSDF) facilities, ozone level in air, PM2.5 level in air.
Location and temperature[58]	Latitude and longitude, average daily temperature (July).
Building characteristics[59]	A/C type in home (4 categorical variables: central A/C, heat pump, room A/C, No A/C), type of home (5 categorical variables; mobile home, single family attached, single family detached, multifamily 2–4 units, multifamily 5+ unit), number of bedrooms in home (5 categorical variables: 1–5 bedrooms).
Power outages[60]	Number of customers impacted, power outage duration (minutes), average outage occurrence.
Low income energy affordability[67]	Energy burden $\left(\frac{\text{energy bill}}{\text{income}}\right)$

Data groups and their respective data sources are created and described to facilitate a clear discussion of the impact of energy burden indicators. Here, the data group represents the label it will be discussed as, and the data descriptions provide the indicators in each group.

indicator importance in machine learning models,[48,61–63] this study uniquely considers a large set of variables (42 indicators) and reduces the number of variables to support immediate climate action for frontline communities, highlighting the most significant indicators as a first step in creating knowledge-informed policies, creating more value and purpose. Further, the combination of data groups offers novel insights into indicators of energy burden beyond socioeconomic and building infrastructure within the U.S.

As the U.S. faces a growing dependence on electrical energy use, novel approaches to enhance comprehension of energy burden are a pressing issue, particularly for frontline communities facing the consequences of a changing climate. A changing climate creates multi-faceted issues, with one challenge being the increased need for A/C in residential spaces.[64] Increased need for A/C is evident when considering weather events, such as the 2022 California heatwave, which led to the California Independent System Operator setting a new peak demand. The Western Interconnection also set a new peak demand in July 2024.[65] However, previous studies have often concentrated on the interconnections of heating energy consumption, poverty, and residential dwellings. Although findings from Wang and Chen[66] suggest that many areas will experience an increase in A/C expenditures but a decrease in heating expenditures as year-round temperatures are projected to increase. Since cooling seasons have been less studied in this context, this study focuses on the summer season, using summer temperature data and investigating A/C types in residential buildings for each U.S. census region regarding county-level energy burden. Each census region and the county-level energy burden within each region are shown in Figure 1.

Overall, this research contributes to the growing knowledge surrounding energy equity using novel data modeling techniques applied to energy burden, specifically, the post hoc interpretations, which can be used to aid in knowledge-informed policy. The main contributions of this work include.

(1) It is one of the first studies to include environmental justice indicators, as defined by the Environmental Protection Agency, and resiliency measures for predicting energy burden. Climate change is projected to impact frontline communities disproportionately; thus, understanding the implications of energy burden beyond energy use, social demographics, and income is crucial to intergenerational equity.
(2) This is one of the first studies dedicated to the feature, or indicator, importance, and influence in terms of predicting energy burden regarding a large dataset of 42 variables and creating smaller subsets for prompt policy action within the U.S.
(3) The development of a data-driven framework to identify indicators of energy burden that is flexible to new inputs and could be used with different ML techniques or geographic scales.

For a better understanding of indicators for energy burden in the U.S., two modeling frameworks are developed and

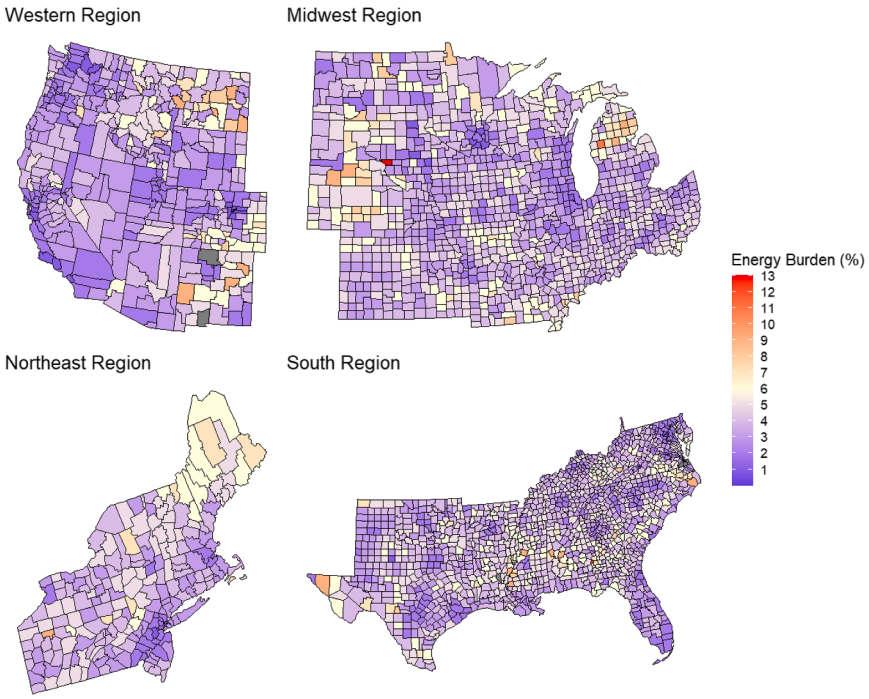


Figure 1. County level energy burden per census region

The percentage of energy burden presented in this figure uses data from the Llow-income energy affordability data tool[67] and is shown at the county level for each census region. The energy burden percentage data used in this figure is used for the predictions throughout the study and is described herein. The purple indicates no or low energy burden. In contrast, the white area indicates an energy burden of 6%, the threshold for a high energy burden, continued by the red areas, which are experiencing a severe energy of 10% or higher.

compared. For both modeling frameworks, the data are subjected to an exhaustive subset selection (ESS) model, which uses a "Branch and Bound" algorithm to determine the optimal 15 indicators for predicting energy burden. From the best 15 predictors, datasets featuring five indicators are created, using all possible combinations of the 15 predictors (this results in 3,003 datasets). However, for modeling framework one, 20 different generalized additive models (GAMs) are developed (60,060 models), and then the 100 best models based on the generalized cross-validation (GCV) score are selected. Each of the model equations for the GAMs is provided in Methods S2, Table S1. For modeling framework two, random forest models are created for each of the 3,003 datasets, and the top 100 models based on the R^2 value are selected. A representative decision tree is created from the random forest model with the highest R^2 value for model interoperability. Both GAMs and random forests are used for their abilities to deal with non-linear and non-monotonic relationships between the indicators and the response variable. GAMs offer superior prediction capabilities and computational efficiency as they are less complex than other methods, such as random forests. However, random forests are used, given their underlying bagging properties and abilities to handle data with high variance. Additionally, as previously mentioned, random forests have been effective in studies regarding climate, energy, and sustainability in past studies. However, Abolafia-Rosenzweig et al.[68] find GAMs to outperform random forests and support vector machines. GAMs have been used in various studies and fields, including studying lead exposure in children,[69] income inequality in healthcare,[70] renewable energy power production,[71] and climate relations to air quality and health.[72] Thus, GAMs show promise in predicting and understanding the complex interactions between the data groups and energy burden. For both modeling frameworks, a leave-one-column-out (LOCO) analysis is completed on each of the top 100 models to understand the indicator influence. Lastly, all models are evaluated for model fit, examples being the R^2 value, root-mean-square error (RMSE), and mean absolute error (MAE). Two datasets were used for each framework: the energy burden full indicator set, which represents the original data from each independent dataset, see Table 1, combined at the county level and standardized. The second dataset is the principal components (PCs). The PCs are taken from a principal-component analysis (PCA) completed for dimensionality reduction. The STAR Methods section and the supplemental information Methods S2 provide a more detailed description of the modeling frameworks. An overview of the modeling frameworks is provided in Figure 2. Overall, this study aims to answer the question "what subset of indicators should be prioritized when creating knowledge-informed policy to alleviate energy burden for each census region in the U.S.?" Assumptions include that the states within each census region (west, midwest, south, northeast) will have similar indicators. The primary limitations include spatial resolution (county-level data) and general data availability, which resulted in using county averages, further discussed in the "limitations of the study" section.

RESULTS

The results presented are divided into two groups corresponding to the datasets used: the full indicator set or the PCs. Indicator groups are discussed to foster a more direct discussion around the indicators of energy burden used in this study. Indicator groups are provided in Table 1.

Energy burden full indicator set feature selection

In order to gain a better understanding of indicator importance for knowledge-informed policy, the indicators selected in the top 100 models for the GAMs are provided in Figure 3; the number of low-income individuals occurred in all of the top 100 models for each region. In comparison, the median age of individuals and poverty status occurred in the top 100 models for

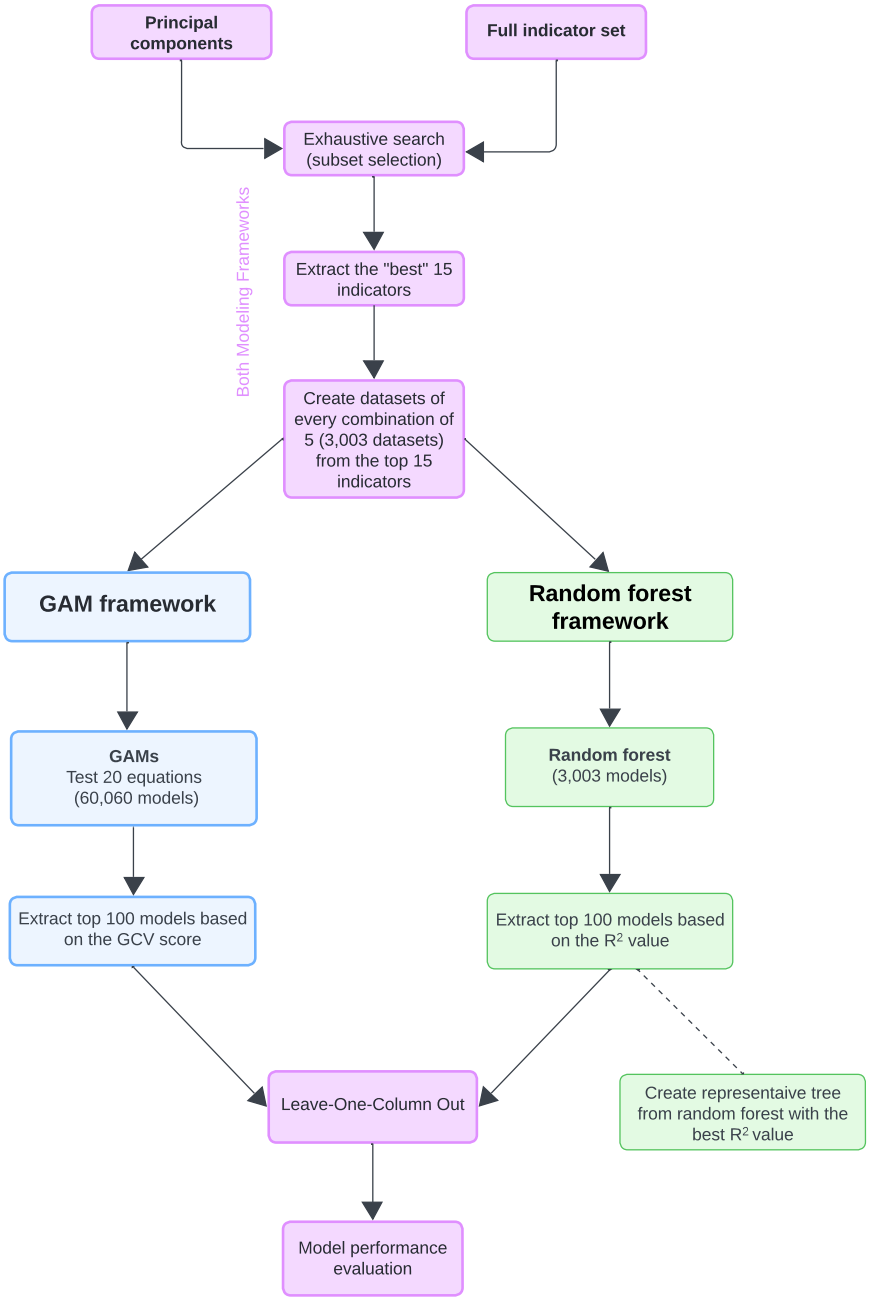


Figure 2. Overview of modeling frameworks

A brief description of the processes taken for each modeling framework. The pink represents processes applied to both modeling frameworks, while the blue is specific to framework one and the green is specific to framework two. Overall, two datasets, the full indicator set, and the principle components, are subjected to an exhaustive subset selection search. Then, 3,003 datasets are created from all combinations of 5 indicators. The datasets are then used in the GAM framework and random forest framework.

the West and Northeast. Additionally, population density occurred in each of the top 100 models for the Midwest. The West and Northeast share similar characteristics, while the South and Midwest have the most variability. For the random forest models, none of the indicators occurred in the top 100 models for every region. Thus, the random forest models are more distributed among the indicators than the GAMs. However, for both the random forests and GAMs, the Midwest shows similarities, as low-income individuals and population density are present in the top 100 models. One main takeaway is that with the ESS, the GAMs and random forests used the same 26 indicators out of the original 42 in the top 100 models. This is significant, as it provides a more tangible set of parameters to inform policy, reducing the parameter size by 16. The results of the ESS are provided in the STAR Methods.

Energy burden principal components feature magnitude

Although the PCA eliminates an aspect of model interoperability, both of the modeling frameworks use PCs to test the predictability of energy burden with less information loss. For the PCA, all 42 variables described in the "data description" from Table 1 are used. To better understand the indicator's contribution to each PC, the absolute value of the magnitude of influence for each data group is provided in Table 2. The magnitude represents the influence indicators have within each PC. The first three PCs for each region were selected, given that this is when the variance explained by each subsequent PC began to drop off. The fraction of the variance explained for the first 20 PCs is provided in S1: Data, Figure S1, and the magnitude of influence is shown for the first 15 PCs in S1: Data Figure S2, accompanied by S1:Data Table S1. The raw values from the PCA are provided in S1:Data Table S2 for the Midwest, S1:Data Table S3 for the Northeast,S1:Data Table S4 for the South, and S1:Data Table S5 for the West.

For each region, the building characteristics had the most significant influence in the first PC, which is the most important PC as it will explain the greatest percentage of the data variance. Following the building characteristics are the socioeconomic variables and environmental justice variables. Among each region, within the first three PCs, the building characteristics, the socioeconomic variables, and the environmental justice variables are the primary influencers, meaning these data groups are represented more than their counterparts within the first three PCs.

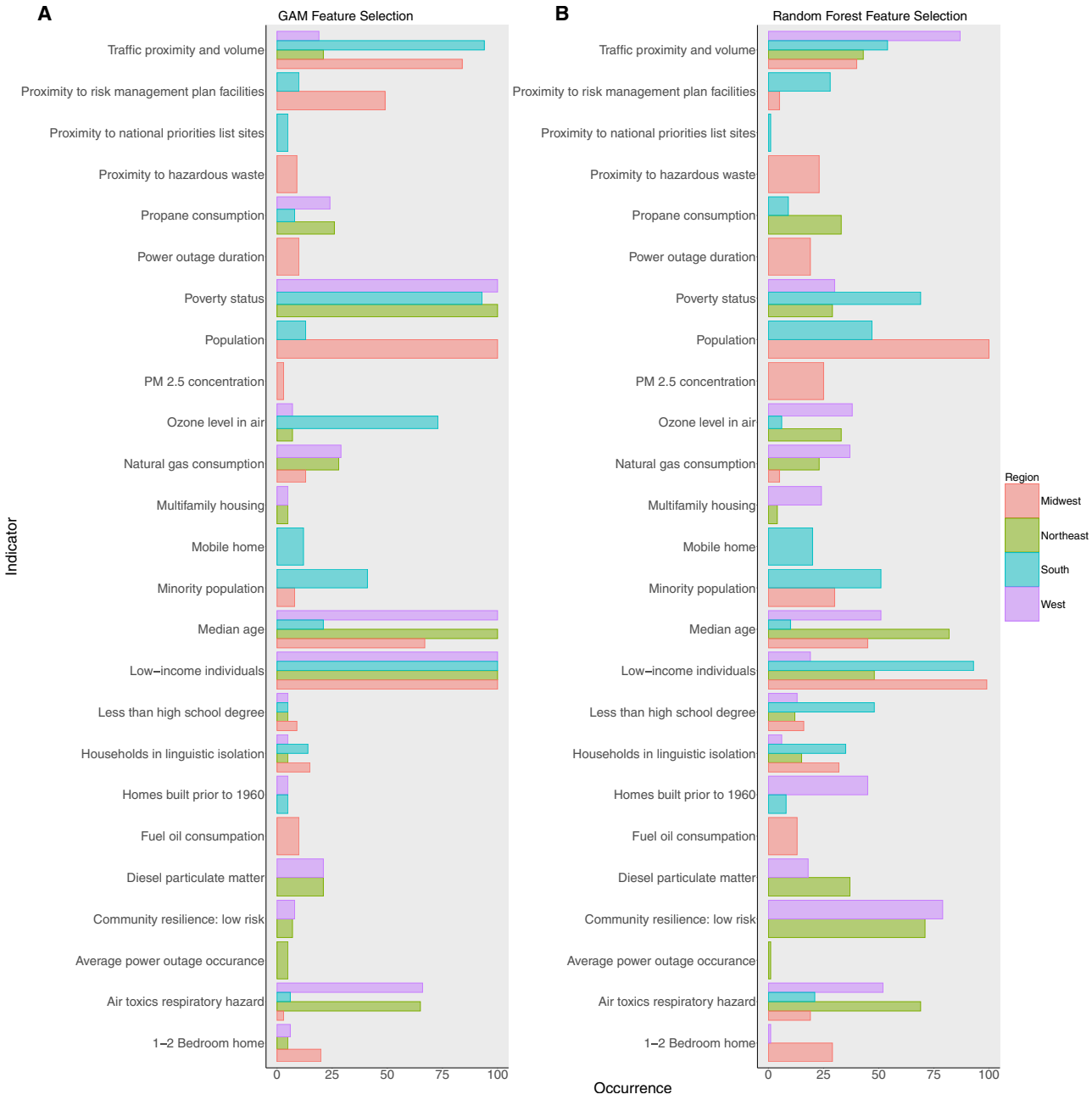


Figure 3. Generalized additive model variable selection for the top 100 models

The variables that appeared in the top 100 models for each region using the full indicator set. Each region is assigned a color, while the bar width depends on the number of regions that include the variable. Each variable is given the same spatial width, but the individual bar size depends on the regions included. For instance, if all regions select the variable in the top 100 models, the bars will be the smallest; if only one region selects the variable, the bar for that region will be the largest. Low-income individuals are included in all of the top 100 models for each region, while median age and poverty status are included in all top 100 models for the West and Northeast.

Energy burden leave-one-column-out analysis

After finding the top 100 models, a LOCO analysis is completed to better understand the indicator influence within the top 100 models for the GAMs and random forests. A LOCO analysis is a type of meta-analysis, meaning the results of multiple independent studies are investigated to determine overall trends. In this case, each of the top 100 models for the GAMs and random forests are recreated, except one column, or indicator, is dropped from each model. The evaluation criteria, such as R^2, RMSE, and MAE, are logged for each model, which drops one value and is compared to the original models, including all model variables. In general, LOCO analyses are performed to determine the

Table 2. Principle-component analysis magnitude

		Magnitude						
Region	PC	Community resilience	Socioeconomic variables	Environmental justice variables	Location and temperature	Building characteristics	Power outages	Total variance explained
Midwest	1	0.67	0.64	0.99	0.01	3.11	0.23	39%
	2	0.27	2.1	1.16	0.14	1.42	0.31	15%
	3	0.03	1.49	1.74	0.35	0.89	0.11	6%
Northeast	1	0.66	0.7	1.04	0.07	3.04	0.23	40%
	2	0.28	2.33	0.82	0.07	1.49	0.28	15%
	3	0.13	1.94	1.57	0.27	0.85	0.33	5%
South	1	0.69	0.55	0.82	0.03	3.11	0.25	38%
	2	0.14	2.52	0.79	0.06	1.15	0.19	12%
	3	0.03	1.18	1.84	0.19	0.82	0.08	7%
West	1	0.66	0.65	1.01	0.06	3.02	0.34	41%
	2	0.26	2.37	0.86	0.08	1.36	0.28	14%
	3	0.12	2	1.54	0.28	0.88	0.16	5%

For each region, the data groups are shown with their respective magnitude of influence within the first three PCs. The building characteristics, followed by the environmental justice variables, consistently show the most significant magnitude of influence within the first PC, which is the PC that explains the highest level of data variance.

overall effect size and influence of individual parameters in a statistical model.[73] Here, the results of the LOCO are provided for each region in Figures 4, 5, 6, and 7, regarding the impacts on the R^2 value.

For additional model evaluation metrics, see the Evaluation Criteria section of the STAR Methods The PC RMSE is found in Figure 8, and the results for the PC MAE are found in Figure 9. For the indicators that show the most significant decrease in the R^2 value, the associated p value is provided to check for the significance in relation to energy burden.

Beginning with the Midwest, in Figure 4, for both the GAM and random forest comparisons for the full indicator set, low-income individuals and population impact the R^2 value the most when left out of the original model. Both of these indicators are additionally included in each of the top 100 models for this region. The population per county being significantly $(p < 0.01)$ correlated to energy burden, this is due to less populated areas having a higher energy burden. However, the count of low-income individuals does not show the same level of significance. This could be due to the energy burden data being averages per county. However, the count of low-income individuals does show significance with indicators such as PM 2.5 concentration and less than a high school education $(p < 0.01)$, which have been linked to socioeconomic disparities in the past.[30] For the PCs, the first PC results in the most significant difference in R^2 value and explains 39% of the data variance. The building characteristics and the environmental justice variables have the highest influence.

In the Northeast, in Figure 5, poverty status $(p < 0.01)$ and low-income individuals $(p < 0.01)$ show the most significant decrease in R^2 for the GAMs. In contrast, community resilience: low risk $(p < 0.01)$ shows the most significant reduction in R^2 for the random forests. It is important to note that the data do contain counties with and without a high energy burden. When limiting the data to only the counties with an energy burden greater than 6%, the p value for community resilience: low risk is no longer significant. The PCs show that the first PC has the most significant impact on the R^2, which accounts for 40% of the data variance and is primarily composed of the building characteristic indicators, followed by the environmental justice indicators.

In the South region, in Figure 6, poverty status $(p < 0.01)$, population $(p < 0.01)$, and low-income individuals $(p < 0.01)$ showed the greatest change in the R^2 values for both GAMs and random forest models for the full indicator set, with lower population, and higher levels of poverty status and low-income individuals being related to energy burden. Regarding the PCs, the first PC had the greatest influence on the R^2 and explained 38% of the data variance. The building characteristics and environmental justice indicators have the largest magnitude of influence.

Lastly, the results for the West region, provided in Figure 7, show that the most significant difference in R^2 values for the GAM models is attributed to poverty status $(p < 0.01)$. At the same time, the random forest models indicate that community resilience: low risk $(p < 0.01)$ results in the greatest decrease in R^2 value. Similar to the other region's PC results, the West region shows that the first PC results in the most significant drop in R^2 value. The first PC explains 41% of the data, and the building characteristics and environmental justice indicators have the greatest magnitude of influence. The consistent effect of the first PC and the building characteristics, together with environmental justice indicators having the most significant influence, suggests that improving building efficiency and addressing environmental justice issues could reduce energy burden in low-income areas.

Model comparison

This section provides an overview of the best model, according to the R^2 value for modeling frameworks one and two, and for the energy burden full indicator set, found in Table 3 and the PCs, found in Table 4. Overall, the models using the PCs outperformed those using the full indicator set. This is to be expected

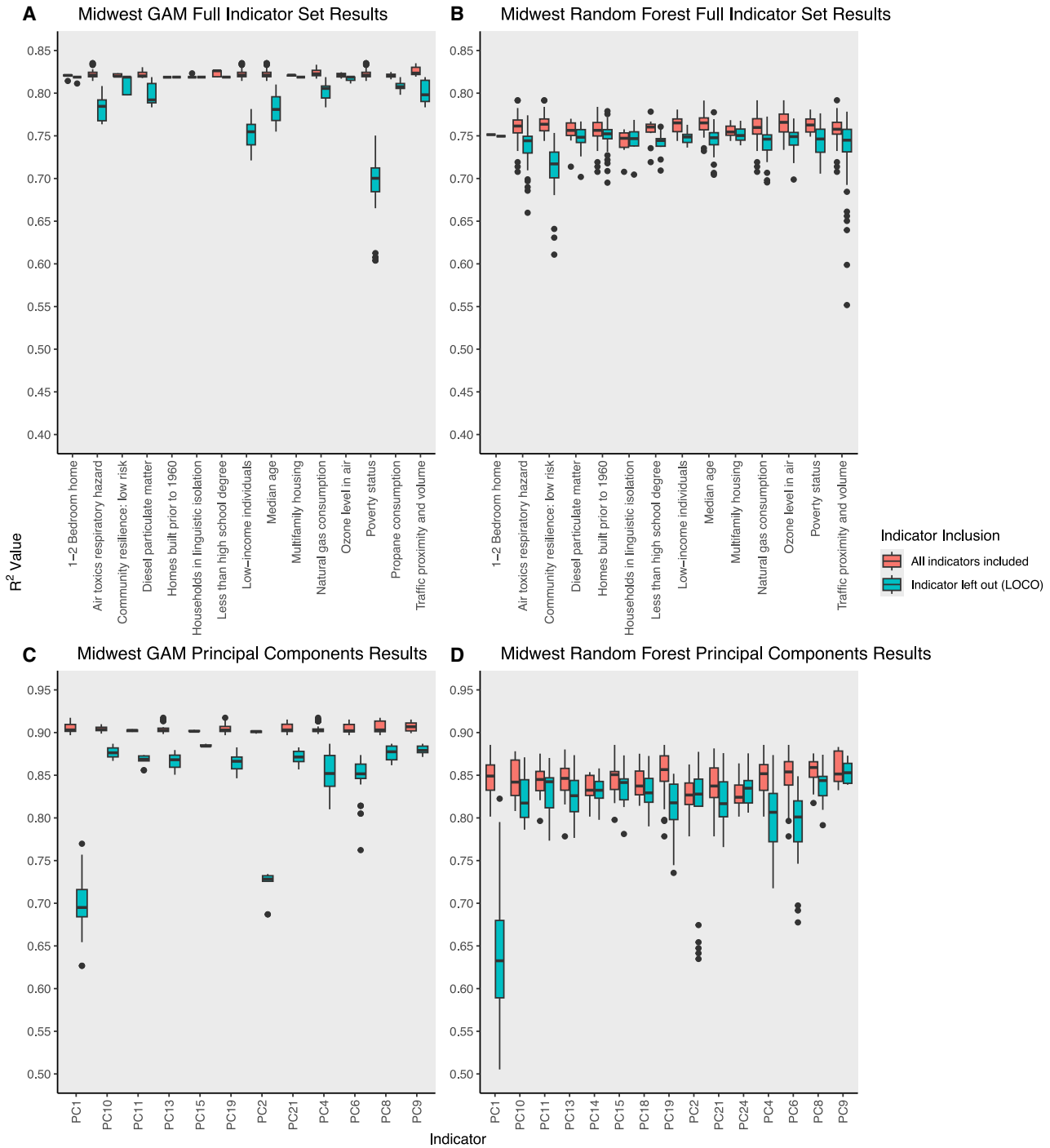


Figure 4. Leave-one-column-out analysis indicator influence for the Midwest region

(A) Represents the LOCO analysis for the GAM full indicator set in the Midwest region.

(B) Represents the LOCO analysis for the random forest full indicator set in the Midwest region.

(C) Represents the LOCO analysis for the GAM PCs in the Midwest region.

(D) Represents the LOCO analysis for the random forest PCs in the Midwest region. Each indicator in the top 100 models for the GAMs and the random forests is subjected to a LOCO analysis. Here, the R^2 value is provided for the entire model, meaning all the indicators for that model were included (red), and the LOCO model, meaning that one indicator was left out of the model (blue). The indicator dropped in the LOCO model is provided by the *x* axis, and the change in R^2 is presented on the *y* axis. Low-income individuals and population show the most significant difference in R^2 value when excluded from the models for the full indicator set. Regarding the PCs, the first PC has the most significant effect on model performance for both the GAMs and random forests.

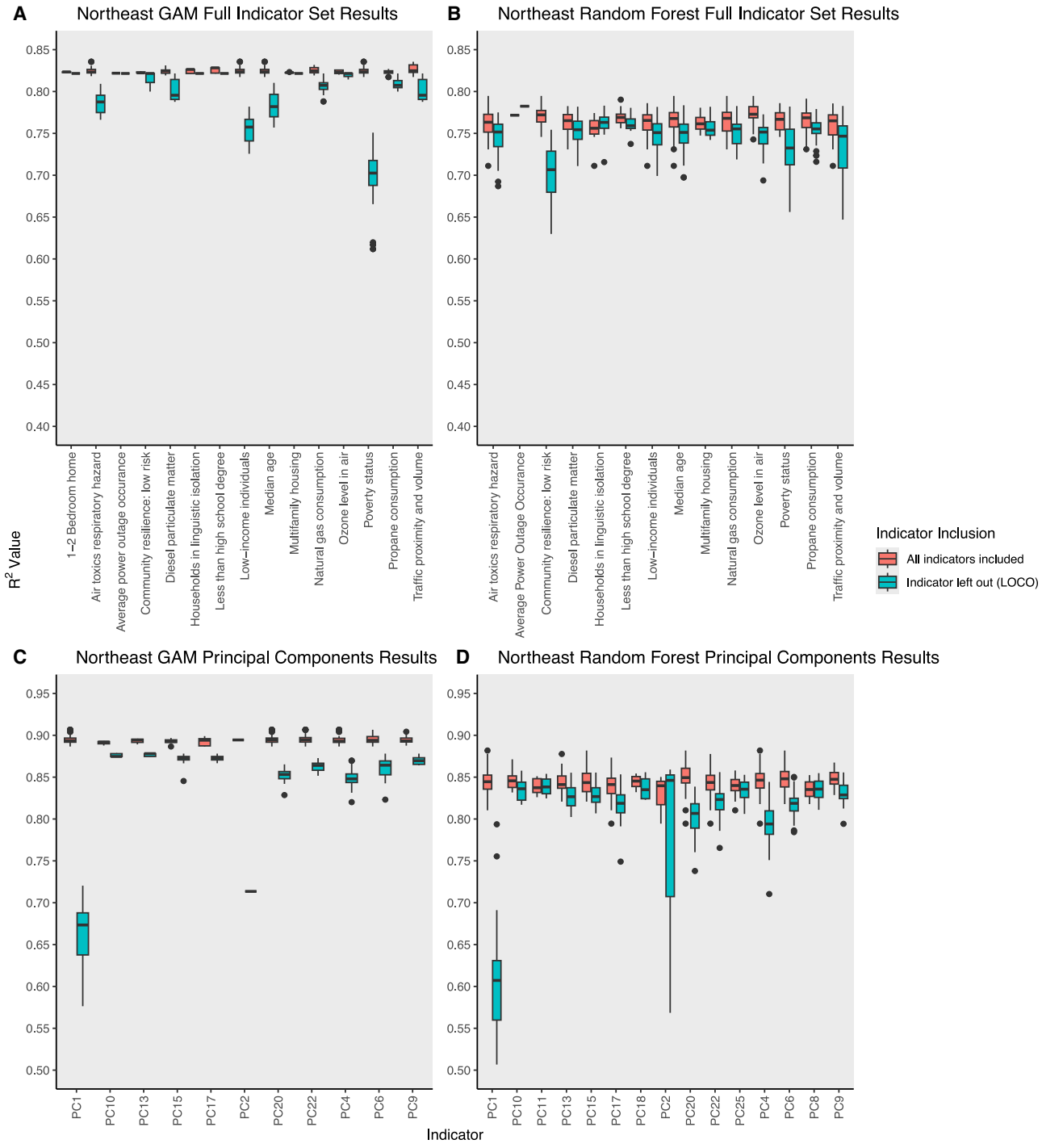


Figure 5. Leave-one-column-out analysis indicator influence for the Northeast region

(A) Represents the LOCO analysis for the GAM full indicator set in the Northeast region.

(B) Represents the LOCO analysis for the random forest full indicator set in the Northeast region.

(C) Represents the LOCO analysis for the GAM PCs in the Northeast region.

(D) Represents the LOCO analysis for the random forest PCs in the Northeast region. Each indicator in the top 100 models for the GAMs and the random forests is subjected to a LOCO analysis. Here, the R^2 value is provided for the entire model, meaning all the indicators for that model were included (red), and the LOCO model, meaning that one indicator was left out of the model (blue). The indicator dropped in the LOCO model is provided by the *x* axis, and the change in R^2 is presented on the *y* axis. For the full indicator set, the GAMs, low-income individuals, have the most significant impact, while community resilience has the most significant impact for the random forests. Regarding the PCs, the first PC has the most significant effect on model performance for both the GAMs and random forests.

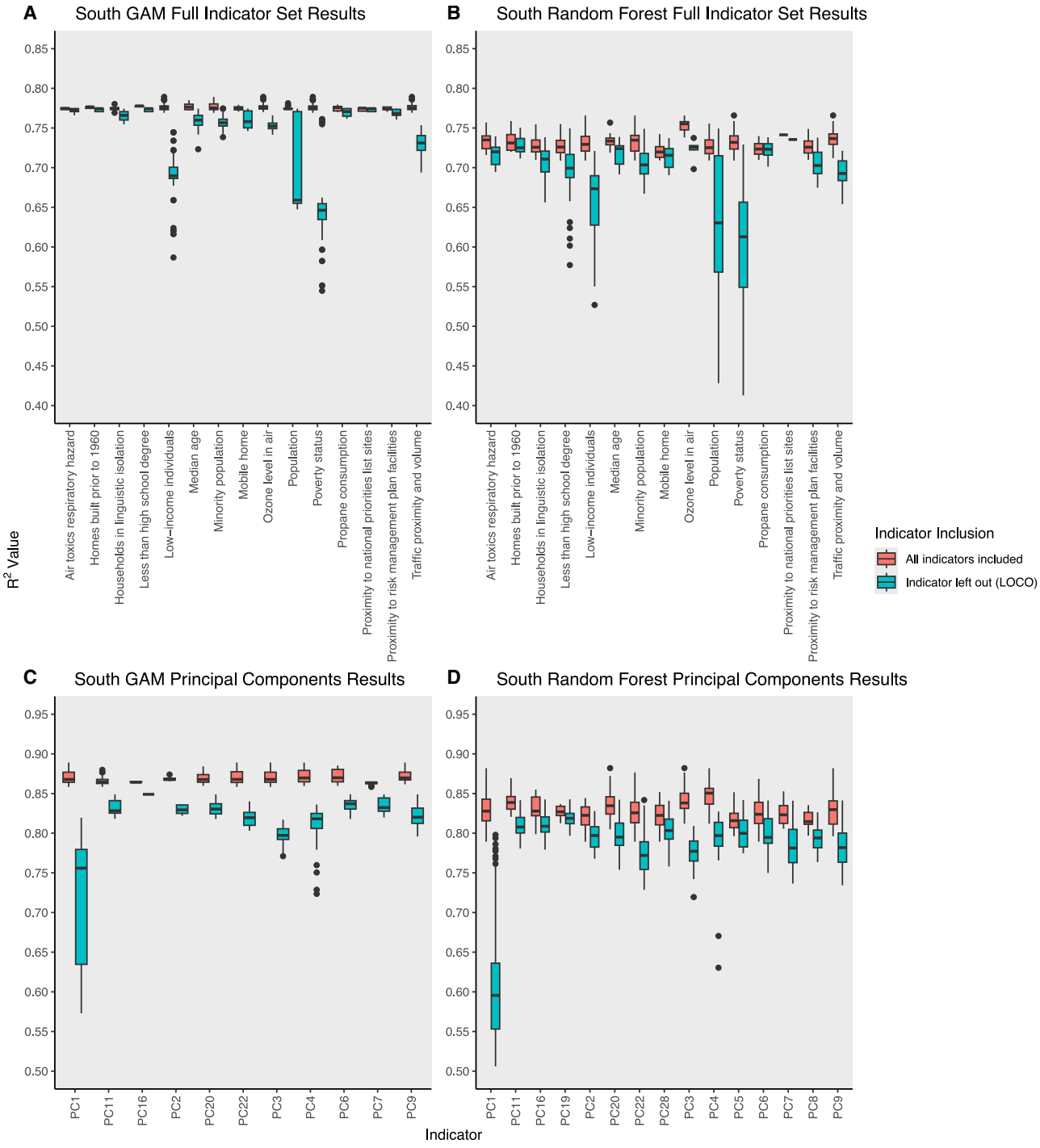


Figure 6. Leave-one-column-out analysis indicator influence for the South region

(A) Represents the LOCO analysis for the GAM full indicator set in the South region.

(B) Represents the LOCO analysis for the random forest full indicator set in the South region.

(C) Represents the LOCO analysis for the GAM PCs in the South region.

(D) Represents the LOCO analysis for the random forest PCs in the South region. Each indicator in the top 100 models for the GAMs and the random forests is subjected to a LOCO analysis. Here, the R^2 value is provided for the entire model, meaning all the indicators for that model were included (red), and the LOCO model, meaning that one indicator was left out of the model (blue). The indicator dropped in the LOCO model is provided by the *x* axis, and the change in R^2 is presented on the *y* axis. Poverty status, population, and low-income individuals significantly impact the R^2 value in both the GAMs and random forests for the full indicator set. Regarding the PCs, the first PC has the most significant effect on model performance for both the GAMs and random forests.

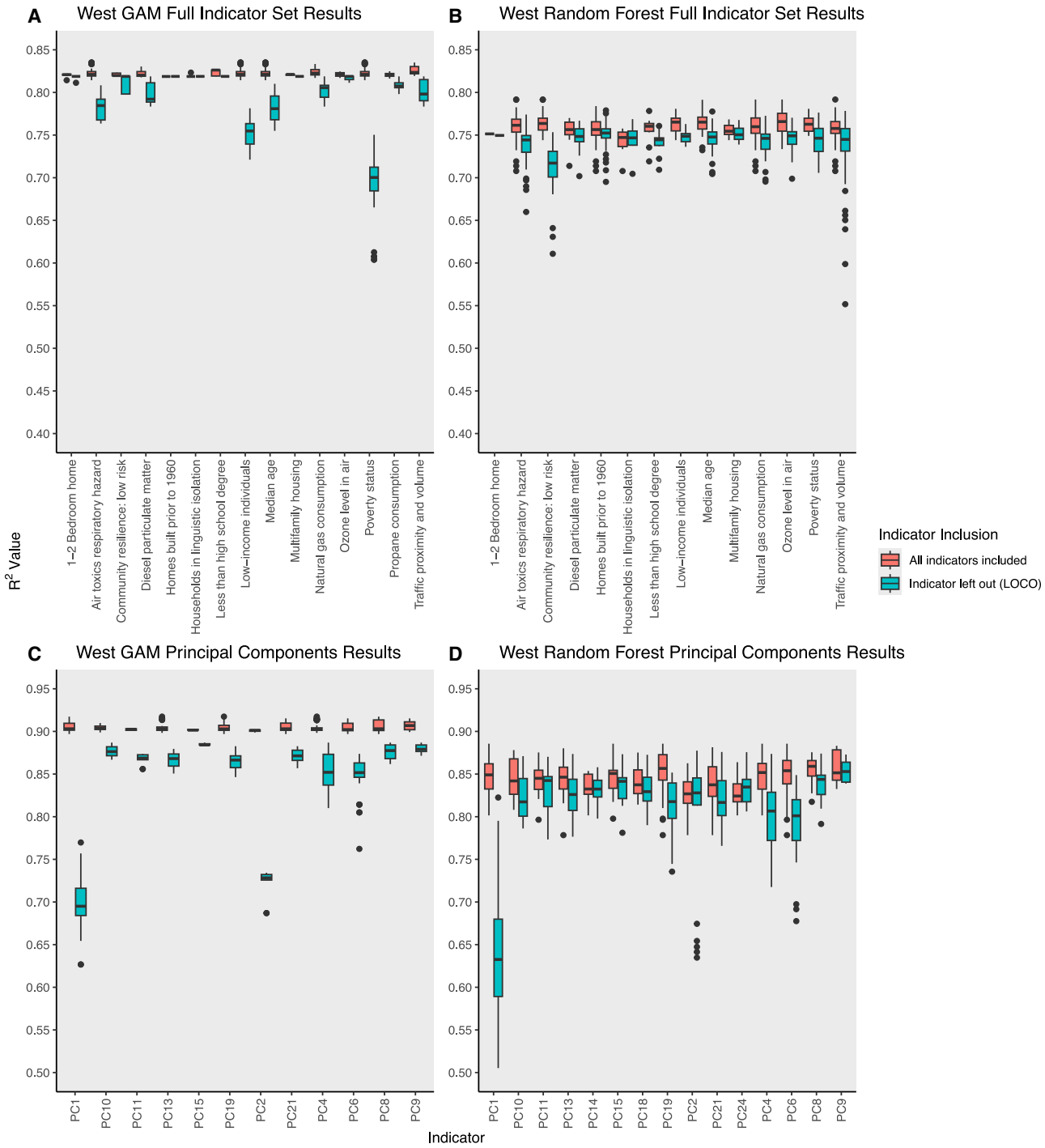


Figure 7. leave-one-column-out analysis indicator influence for the West region

(A) Represents the LOCO analysis for the GAM full indicator set in the South region.

(B) Represents the LOCO analysis for the random forest full indicator set in the South region.

(C) Represents the LOCO analysis for the GAM PCs in the South region.

(D) Represents the LOCO analysis for the random forest PCs in the South region. Each indicator in the top 100 models for the GAMs and the random forest is subjected to a LOCO analysis. Here, the R^2 value is provided for the entire model, meaning all the indicators for that model were included (red), and the LOCO model, meaning that one indicator was left out of the model (blue). The indicator dropped in the LOCO model is provided by the *x* axis, and the change in R^2 is

(legend continued on next page)

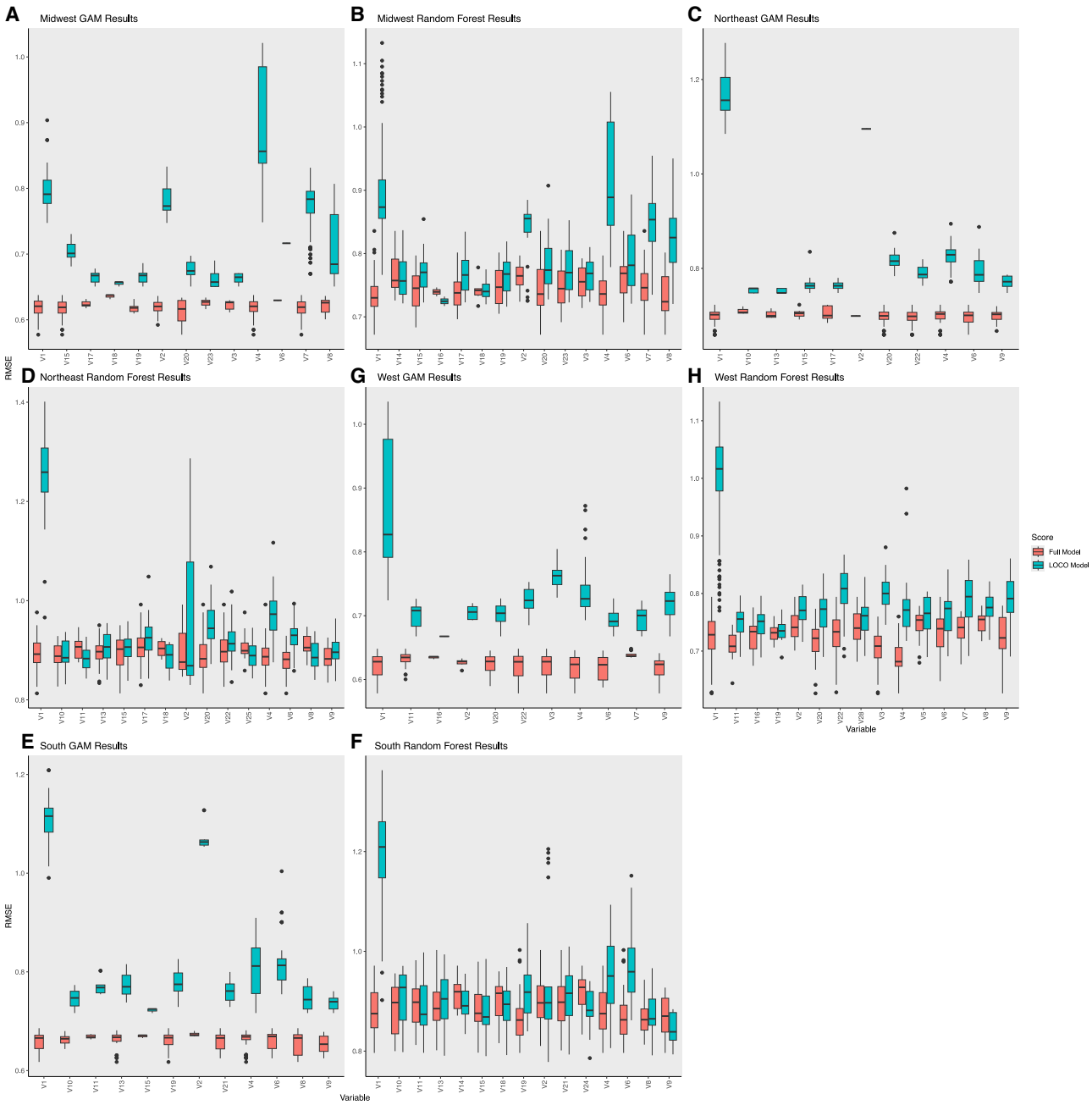


Figure 8. Principal component leave-one-column-out root-mean-square error

as the PCA creates linear combinations of the original variables; thus, greater variance is captured using 5 PCs rather than using 5 indicators independently. For both the full indicator set and PCs, the GAM models outperformed the random forest models in terms of the R^2 value, except for in the West region, using the full indicator set, which has the same R^2 value of 0.84. For the full indicator set, the MAE, which does not take into account the direction of the error, and RMSE show mixed results. In the Midwest and the West, the MAE is lower for the GAM, but for the Northeast and South, the random forest outperforms the

presented on the *y* axis. Poverty status, population, and low-income individuals have the most significant impact on the R^2 value in both the GAMs and random forests for the full indicator set. Regarding the PCs, the first PC has the most significant effect on model performance for both the GAMs and random forests. For the GAMs, poverty status has the most significant impact, while traffic proximity and volume have the most significant impact for the random forests for the full indicator set. Regarding the PCs, the first PC has the most significant effect on model performance for both the GAMs and random forests.

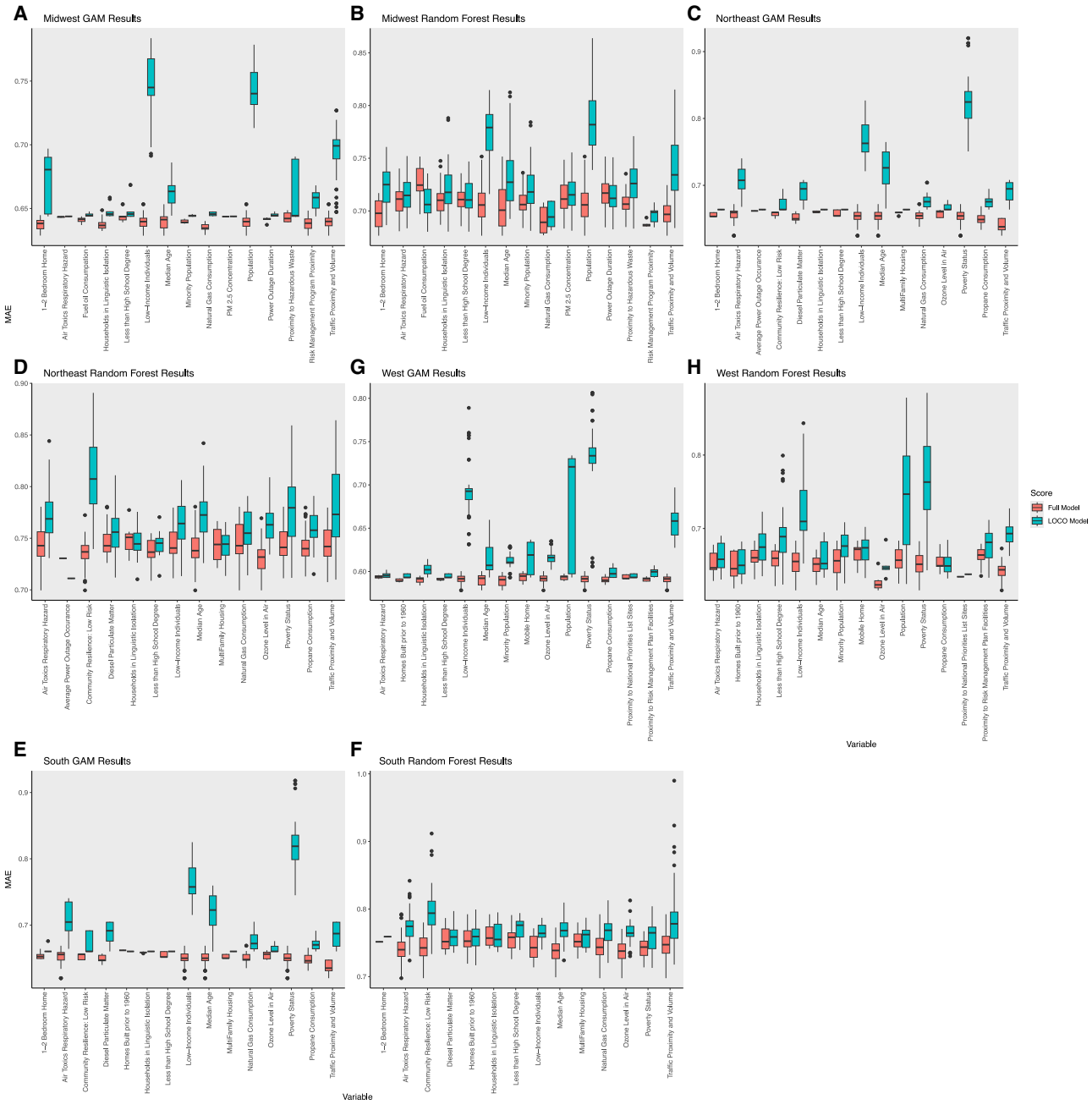


Figure 9. Principal component leave-one-column-out mean absolute error

GAMs. However, the random forest models outperform the GAMs in terms of the RMSE scores for all regions except the Midwest, where the GAM outperforms the random forest. For the PCs, the GAM models consistently have better MAE and RMSE values.

Random forests and GAMs are both additive in nature and handle non-linear relationships within the data. However, there are key differences that could result in the differences between model performance and feature selection. In the GAMs, the linear components are replaced with a smooth non-linear function. The additive component occurs as each individual, smoothed non-linear indicator is added together to create an estimate. In the random forest model, predictions are combined or added from a sequence of models (decision trees) to create a prediction. Additionally, random forests use bagging, meaning each of the decision trees is trained on a subset of the data. As part of the bagging, indicators that contribute more heavily to the prediction are often selected over those that contribute to noise or are insignificant. GAMs do not use bagging; rather, they use spline methods, typically smoothing splines. In their

Table 3. Energy burden full indicator set best models

Region	Model	Representative model indicators	R^2	RMSE (%)	MAE (%)
Midwest	GAM	Total population, low-income individuals, traffic proximity, proximity to risk management plan facilities, median age	0.77	0.87	0.63
	Random forest	Total population, low-income individuals, traffic proximity, pm 2.5 concentration, median age	0.73	0.93	0.68
Northeast	GAM	Poverty status, low-income individuals, air toxics respiratory hazard, traffic proximity, median age	0.84	0.86	0.62
	Random forest	Low-income individuals, air toxics respiratory hazard, ozone level in air, community resilience: low risk, median age	0.81	0.57	0.45
South	GAM	Poverty status, minority population, low-income individuals, ozone level in air, traffic proximity	0.78	0.77	0.58
	Random forest	Poverty status, minority population, low-income individuals, linguistically isolated, traffic proximity	0.72	0.67	0.51
West	GAM	Poverty status, low-income individuals, air toxics respiratory hazard, traffic proximity, median age	0.84	0.85	0.62
	Random forest	Homes built prior to 1960, air toxics respiratory hazard, traffic proximity, community resilience: low risk, median age	0.84	0.81	0.64

The model with the highest R^2 value for each region and modeling framework using the full indicator set. The region, modeling method, and the indicators used are shown, with the corresponding R^2, RMSE, and MAE values. In terms of R^2 the GAMs outperform the random forests in all regions except for the West region, which has the same R^2 value. Regarding the MAE and RMSE, the best performing models show mixed results among the GAMs and random forests. Given the different modeling methods, indicators used, and regions, this table should not be used as a direct comparison.

simplest form, smoothing splines estimate the functional relationship between an indicator, such as low income and energy burden. The indicator is then transformed by the functional relationship; this transformed indicator is then used in the prediction. The STAR Methods Section and the supplemental information Methods S2 provide more information about the methods used.

Representative trees

Given that random forests are a black box method that utilizes decision trees, and decision trees are considered a highly interpretable method, representative trees from the best performing random forest are provided in Figures 10, 11, 12, and 13. For the Midwest region, the PM 2.5 concentration and traffic proximity have a high influence on high energy burden nodes. For the Northeast median age, low-income individuals, community resilience, and air toxins are indicators of energy burden. The South region shows that the ozone level in the air and poverty status significantly affect a high energy burden. The West shows similar patterns to the Northeast, with a strong influence from the median age, low-income individuals, and community resilience indicators.

Policy implications

The purpose of this study is not how to create policies to alleviate energy burden or energy poverty but to offer insights into the indicators of energy burden that advocates, and policymakers should prioritize and ways that the outcomes of machine learning models can appropriately be used for knowledge-informed policy. Equity initiatives should prioritize transparency, accountability, and fair and just consideration of socioeconomic targeting.[48] Through transparent communication, assistance programs must gain the public's trust, especially when creating knowledge-informed policies supported by science and machine learning. Thus, presenting the outcomes from the GAMs and representative decision trees from the random forest models could aid in trust and understanding over black box models that do not offer interpretable outcomes. For instance, decision trees provide a visual illustration that can be understood without an intensive machine-learning background. For each region, a policymaker could identify the indicators that lead to a high energy burden and create a policy that alleviates another injustice (poor air quality, lack of renewable energy, upgrading the power infrastructure to rural areas to reduce outages, low-income), or identify the characteristics (racial identity, educational level, age, etc.) to create policies that are inclusive to different identities and cultures, which is further elaborated on in the discussion.

Using the LOCO analysis results could aid a policymaker, advocate, or government entity in determining the influence of indicators and which ones to prioritize. A majority of existing programs focus on financial support for low-income households and energy efficiency programs for residential upgrades. The results from all models highlight low-income or poverty status as an essential indicator of energy burden, which could

Table 4. Energy burden principal component analysis best models

Region	Model	Representative model indicators (PCs)	R^2	RMSE (%)	MAE (%)
Midwest	GAM	1, 4, 7, 15, 20	0.90	0.57	0.44
	Random forest	1, 4, 7, 8, 20	0.87	0.67	0.50
Northeast	GAM	1, 4, 6, 20, 22	0.91	0.66	0.52
	Random forest	1, 4, 6, 15, 20	0.88	0.81	0.61
South	GAM	1, 3, 4, 9, 22	0.89	0.57	0.44
	Random forest	1, 3, 4, 9, 20	0.88	0.62	0.46
West	GAM	1, 4, 8, 13, 19	0.92	0.62	0.48
	Random forest	1, 4, 6, 15, 19	0.89	0.80	0.61

The model with the highest R^2 value for each region and modeling framework using the PCs. The region, modeling method, and the indicators used are shown, with the corresponding R^2, RMSE, and MAE values. For PCs, the GAMs outperformed the random forests for each of the modeling metrics.

aid in the general public's trust that initiatives are investing in the correct types of programs. However, these programs are currently insufficient to achieve an equitable energy transition. Additionally, the PCA found the building characteristics to be one of the most significant data groups throughout each region, which supports the progression of energy efficiency programs for frontline communities.

From an environmental justice perspective, investigating the cause of poor air quality could lead to initiatives to invest in renewable energy, as an air quality metric (ozone, PM 2.5, air toxins related to respiratory hazard) was commonly found among the best models. Replacing high emissions and more expensive (from an operational standpoint) energy sources and replacing them with renewable energy sources (reduce emissions and are less costly to operate) could reduce air pollutants, aid in solving other environmental injustices, including water use and pollution from power sources, and result in lower electricity prices. Highlighted in Scheier and Kittner,[74] investments in renewable energy, especially in DERs, are often not realized by minority and low-income groups. This creates further disparities within the energy transition as rooftop solar could decrease a household's energy bill, decreasing their energy burden, further Spandagos et al.[48] finds oil or gas dependency to contribute to energy poverty as these prices may fluctuate over time and increase when natural disasters occur, which could result in power shutoffs to vulnerable communities due to the inability to pay. Additionally, local governments could use the results to identify areas to reconsider zoning permits or increase the hosting capacity of the power grid, as Brockway et al.[6] found that minority and disadvantaged communities had less access to rooftop solar due to outdated power infrastructure, which could also lead to more frequent and prolonged power outages. The total population and/or traffic proximity is another indicator often found in the top models, indicating rural areas experiencing a high energy burden; a compounded effect is that these communities usually do not live near critical infrastructure, which is traditionally prioritized in power restoration.[34] Age, socioeconomic status, and location impact a household's willingness to evacuate during power outages or natural disasters, with lower-income and minority communities being less likely to evacuate.[75] Working with local communities and identifying areas for power grid investment, consideration of micro-grid strategies, and general aid in preparedness could contribute to integrating social vulnerability and energy poverty into disaster management and create a path for frontline communities to benefit from the clean energy transition directly.[76,77]

DISCUSSION

The modeling frameworks used in this study were selected to better understand indicators for energy burden in the U.S. and to determine general patterns and distributions related to characteristics attributed to energy burden through the LOCO analysis. Understanding energy burden indicators is critical as it could inform policymakers regarding energy-burdened areas and provide insight into geographic regions that could benefit from policy or aid as household energy demand increases. Adverse consequences are placed on households where energy is unaffordable, which is amplified by a changing climate.[4] The drivers of energy burden and the communities that will face adverse effects include households with low-income individuals, communities experiencing poor outdoor air quality, and communities with a greater population of older or younger community members, as found through the full indicator set. The most influential indicators from the PCA include the building characteristics and environmental justice indicators (including air quality metrics). However, the nexus between environmental justice, socioeconomic factors, housing, community resilience, and power outages are not always independently associated, as shown in Figure 14.

Such policies and the knowledge acquired through learning the indicators of energy burden create a path to climate action to avail and eliminate the compounded burdens of climate change on marginalized communities. Current challenges among the energy and climate nexus are intergenerational equity concerns by nature. For instance, a lack of action by governing bodies now regarding climate change will have profound impacts on the way future individuals will live and experience the world. However, the most harmful impacts of climate change will be on impoverished areas, highlighting the confluence between intergenerational and intragenerational equity. Communities or regions that struggle with meeting needs now may not be able to fulfill obligations to future generations.[78] This is another reason it is vital to understand and address the issues surrounding environmental and energy justice in the present.

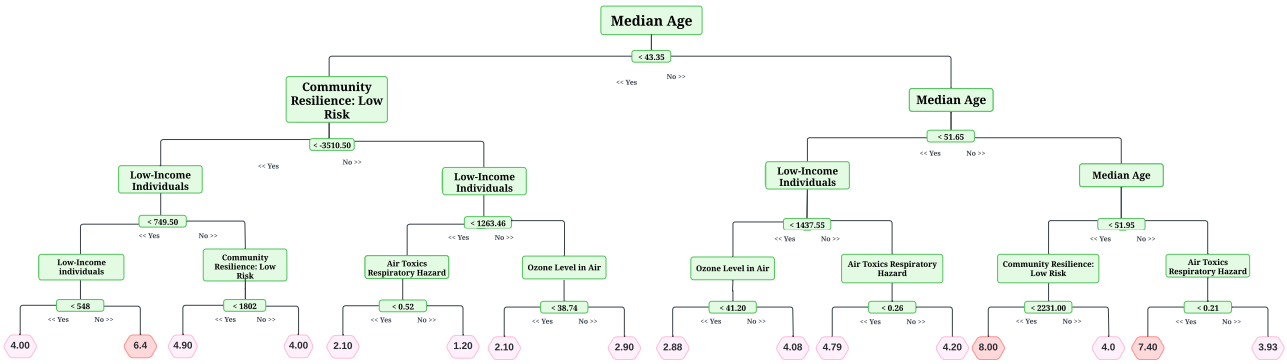


Figure 10. Midwest representative tree

A representative tree from the Midwest random forest model. The high energy burden nodes are in red, while the low energy burden nodes are in pink (in percentile). The root node, or starting point, is the largest to signify importance; each node, indicator split, decreases in size thereafter. For the Midwest, traffic proximity (the root node), PM 2.5, and low income have a large influence on the high energy burden percentages.

From a procedural and distributive equity lens, the modeling methods and feature selection work provided are the first steps in a more extensive process for creating awareness, policies, and programs. This is one of the first studies to investigate energy burden and the intersections among all data groups. Additionally, obtaining the knowledge surrounding the most significant indicators in terms of predicting energy burden offers insights to policymakers. For instance, median age was found in all of the top 100 models for the GAMs in both the West and Northeast. When compared to the South and Midwest, the West and Northeast median age had a higher standard deviation, meaning the median age was more disbursed. However, when looking at the median age of the counties currently experiencing a high energy burden, the West and Northeast have higher average ages but still a higher standard deviation. For the Midwest, the top 100 models for both the GAMs and random forests contain population. When considering the population of the high energy burden counties, it is evident that high energy is associated with rural areas in the Midwest. It is important to note that the respective models focused on selecting the best subsets of data that together best predict energy burden. This is useful in navigating these complex and multi-faceted issues, especially regarding recognitional justice. When looking at the 25 indicators selected by the GAMs and random forests, a policymaker could learn that a low-income individual in a rural area in the Midwest, within a specific age range, may be more likely to experience a high energy burden.

With this information, implementing recognitional and procedural equity could be achieved by holding public forms, with an emphasis on providing the means for vulnerable groups to attend to voice what is viewed as fair and accessible solutions to their community.

On the contrary, selecting the best subsets of data that together best predict energy burden does not mean that other indicators are not important or should not be addressed. For instance, indicators related to air quality have higher values in high energy burden areas but do not always appear in the top 100 models. Overall, this work found that energy burden is a metric that can be predicted with marginal error, especially when using the PCs. Findings conclude that there are marginal differences between the GAMs and the

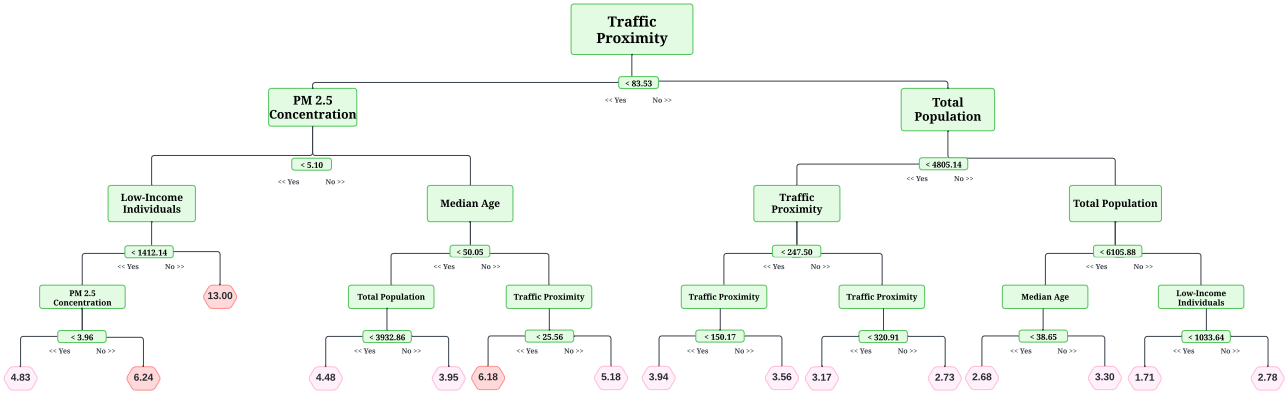


Figure 11. Northeast representative tree

A representative tree from the Northeast random forest model. The high energy burden nodes are in red, while the low energy burden nodes are in pink (in percentile). The root node, or starting point, is the largest to signify importance; each node, indicator split, decreases in size thereafter. The median age (the root node) and low income have a large influence on the high energy burden percentages.

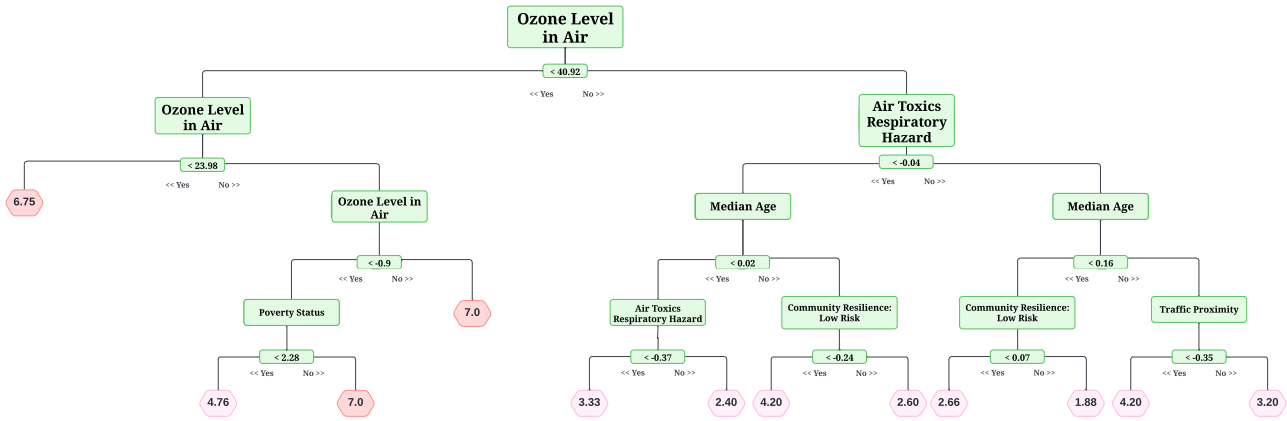


Figure 12. South representative tree

A representative tree from the South random forest model. The high energy burden nodes are in red, while the low energy burden nodes are in pink (in percentile). The root node, or starting point, is the largest to signify importance; each node, indicator split, decreases in size thereafter. The ozone level in the air (the root node) and poverty status have a large influence on the high energy burden percentages.

random forests when using the full indicator set, with the GAMs outperforming the random forests for every region regarding the R^2 value, with mixed results for the RMSE and MSE. When considering the PCs, the GAMs consistently outperformed the random forests in each region and for the R^2, RMSE, and MAE values. Additionally, the models utilizing the PCs outperform the models using the full indicator set.

The PCA identified the building characteristics and environmental justice indicators as the most important throughout the first three PCs. For each region, dropping the first PC in the LOCO analysis resulted in the most significant decrease in the R^2 value for both the GAMs and random forest models. This is to be expected as the first PC explains the highest amount of variance within the data. Thus, building characteristics and environmental justice groups should be used in policy decisions for an equitable energy transition and environmental considerations within a changing climate. As an example, policy incentives to retrofit existing infrastructure, such as older residential buildings, which are predominant in the Northeast, could result in lower energy burdens and a higher quality of life for residents. Considering recognitional equity, socioeconomic factors should influence the type of retrofits, as cultural differences and household identities must be considered within the energy and environmental policies. Additionally, such policies could reduce greenhouse gas emissions, potentially having a positive impact on both household residences and the environment. Understanding the link between energy burden, community resilience, sociodemographics, and building characteristics is essential in creating equitable policies for a diverse demographic region and working closely with communities at a local level, understanding the long and short-term implications, education, and a focus on equitability and affordability. Further, climate change has the potential to disproportionately impact low-income or minority communities.[79] Thus, primitively understanding the characteristics of these communities will aid in climate change adaptation and preparedness.

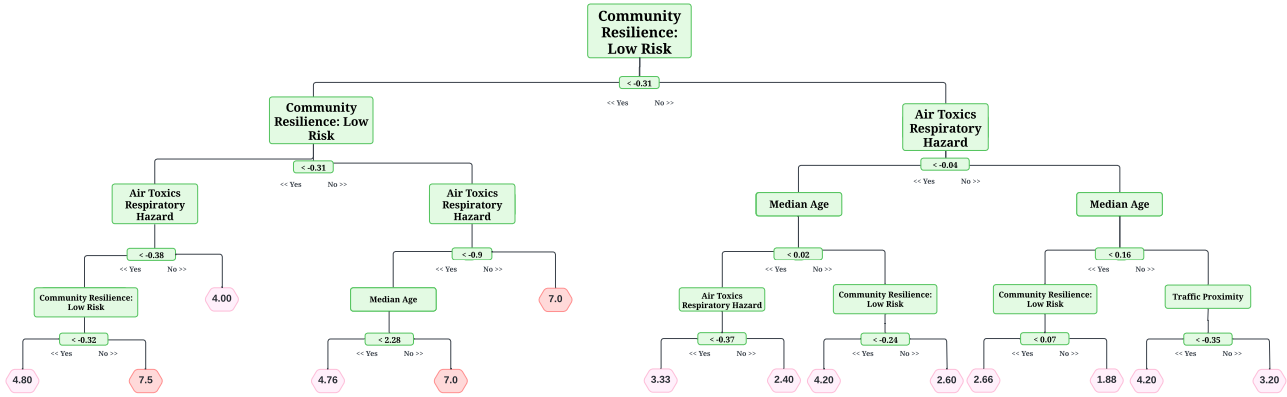


Figure 13. West representative tree

A representative tree from the West random forest model. The high energy burden nodes are in red, while the low energy burden nodes are in pink (in percentile). The root node, or starting point, is the largest to signify importance; each node, indicator split, decreases in size thereafter. Community resilience (the root node), median age, and air toxins related to respiratory risk greatly influence the high energy burden percentages.

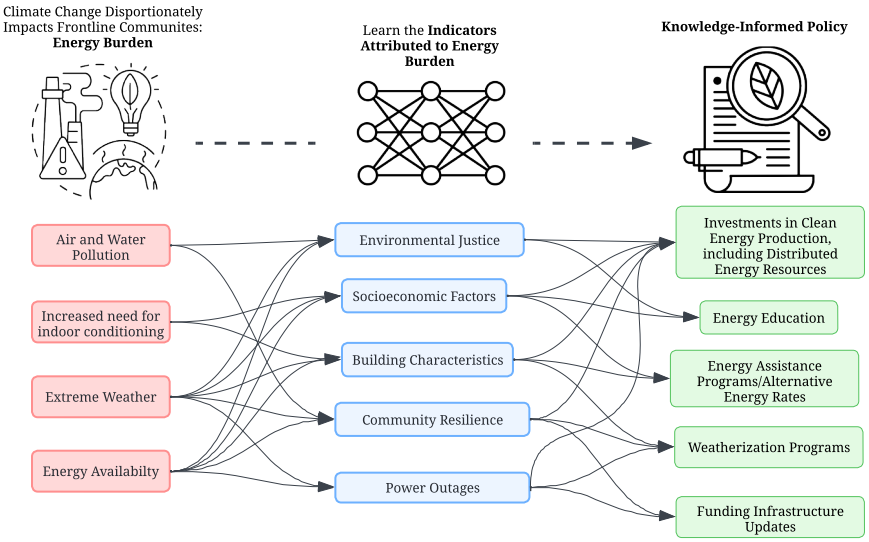


Figure 14. The nexus between data groups and knowledge-informed policy

In red, examples of how climate change impacts marginalized communities are connected to data groups used in this study, shown in blue. Further, the data groups are connected to the types of policies a policymaker could create.

Limitations of the study

To further build upon the models used in this study, the random forest models could be expanded upon by storing the decision trees to create a distribution of energy burden for each county, which could be used as a probability in accessing a county's risk for having a high energy burden. Additionally, spatial-temporal modeling methods should be considered in the future. These include the use of Bayesian spatial models, as they are often helpful in determining dependencies and patterns in space, which has the potential to offer more informed knowledge of energy burden. Previous studies have found low-income U.S. households have less access to updated technologies such as demand response and energy-efficient appliances which could be linked to the age of the home or renter status. Thus, renter status and the age of the home beyond if it was built prior to 1960 could be an additional data group in the future.[80] This study uses county-level data, which could misrepresent small communities in large counties, as such small census regions in counties with high-income census regions in the same county may be overshadowed. It is well documented[81,82] that the spatial resolution and general data availability are challenges in socioeconomic and public health research. This is largely due to privacy concerns, as census tracts are more granular than zip codes, counties, or other geographic regions provided by the Census Bureau of the U.S. Some data sources used in this study, such as the average energy burden[67] are publicly available at the census tract level. However, to achieve this level of spatial resolution, they are subjected to interpolation or resampling, which can often create bias in the data.[83] However, other data sources, such as the building characteristics, were not available at the census tract level.[84] This creates a need for secure, publicly available data at the census region level for the socioeconomic factors.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Jasmine Garland (jasmine.garland@colorado.edu).

Materials availability

- This study did not generate new datasets or reagents.

Data and code availability

- All data reported in this paper will be shared by the lead contact upon request.
- All code generated for this study will be shared by the lead contact upon request.
- All other items generated for this study will be shared by the lead contact upon request.

ACKNOWLEDGMENTS

This work was funded by the U.S. Department of Education Graduate Assistantships in Areas of National Need Fellowship and National Science Foundation Graduate Research Fellowship.

AUTHOR CONTRIBUTIONS

J.G. completed the coding (data manipulation, modeling, and visualizations), writing, and aided in methods development; K.B. and B.L. aided in writing and reviewing the paper, and methods development; R.B, aided in methods development and conceptualization.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Data acquisition and processing
  - Principle component analysis
  - Data modeling
  - Stepwise subset selection
  - Generalized additive model
  - Random forest
  - Leave-one-column-out
  - Evaluation criteria
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112559.

Received: September 30, 2024
Revised: December 26, 2024
Accepted: April 26, 2025
Published: May 6, 2025

REFERENCES

1. Administration, U. E. I.. What is the united states' share of world energy consumption?. https://www.eia.gov/tools/faqs/faq.php?id=87&t=1(2021/11/26.
2. Brown, M.A., Soni, A., Lapsa, M.V., Southworth, K., and Cox, M. (2020). High energy burden and low-income energy affordability: conclusions from a literature review. Prog. Energy *2*, 042003. https://doi.org/10.1088/2516-1083/abb954. https://www.osti.gov/biblio/1731048.
3. Golubchikov, O., and Deda, P. (2012). Governance, technology, and equity: An integrated policy framework for energy efficient housing. Energy Policy *41*, 733–741. https://doi.org/10.1016/j.enpol.2011.11.039.
4. Jessel, S., Sawyer, S., and Hernández, D. (2019). Energy, Poverty, and Health in Climate Change: A Comprehensive Review of an Emerging Literature. Front. Public Health *7*, 357. https://doi.org/10.3389/fpubh.2019.00357.
5. Dobbins, J., and Tabuchi, H. Texas blackouts hit minority neighborhoods especially hard. Available at https://www.nytimes.com/2021/02/16/climate/texas-blackout-storm-minorities.html(2021/11/26).
6. Brockway, A.M., Conde, J., and Callaway, D. (2021). Inequitable access to distributed energy resources due to grid infrastructure limits in california. Nat. Energy *6*, 892–903. https://doi.org/10.1038/s41560-021-00887-6.
7. Monyei, C.G., Sovacool, B.K., Brown, M.A., Jenkins, K.E.H., Viriri, S., and Li, Y. (2019). Justice, poverty, and electricity decarbonization. Electr. J. *32*, 47–51. https://doi.org/10.1016/j.tej.2019.01.005.
8. (2024). Climate Adaptation Partnerships Program in the Justice40 Initiative. Climate and Societal Interactions Division. National Oceanic and Atmospheric Administration. https://cpo.noaa.gov/divisions-programs/climate-and-societal-interactions/cap-risa/justice40-initiative/.
9. Carley, S., and Konisky, D.M. (2020). The justice and equity implications of the clean energy transition. Nat. Energy *5*, 569–577. https://doi.org/10.1038/s41560-020-0641-6.
10. McCauley, D., Heffron, R., Stephan, H., and Jenkins, K. (2013). Advancing energy justice: The triumvirate of tenets. Int. Energy Law Rev. *32*, 107–110.
11. Weiss, E.B. (1989). Climate change, intergenerational equity and international law: An introductory note. Climatic Change *15*, 327–335. https://doi.org/10.1007/BF00138858.
12. Tarekegne, B., Pennell, B., Preziuso, D., and O'Neil, R. (2021). Review of Energy Equity Metrics. In PNNL-32179, p. 1830804. https://doi.org/10.2172/1830804.
13. Cong, S., Nock, D., Qiu, Y.L., and Xing, B. (2022). Unveiling hidden energy poverty using the energy equity gap. Nat. Commun. *13*, 2456. https://doi.org/10.1038/s41467-022-30146-5.
14. Primc, K., Slabe-Erker, R., and Majcen, B. (2019). Constructing energy poverty profiles for an effective energy policy. Energy Policy *128*, 727–734. https://doi.org/10.1016/j.enpol.2019.01.059.
15. Eisenberg, J. F. Weatherization assistance program technical memorandum background data and statistics on low-income energy use and burdens. Available at https://info.ornl.gov/sites/publications/Files/Pub49042.pdf.
16. Drehobl, A., Ross, L., and Ayala, R. (2020). How High Are Household Energy Burdens? (American Council for an Energy-Efficient Economy).
17. Lewis, J., Hernández, D., and Geronimus, A.T. (2019). Energy efficiency as energy justice: addressing racial inequities through investments in people and places. Energy Effic. *13*, 419–432. https://doi.org/10.1007/S12053-019-09820-Z.
18. Recs: One in three US households faced challenges in paying energy bills in 2015. https://www.eia.gov/consumption/residential/reports/2015/energybills/.
19. Brown, M.A., Soni, A., Doshi, A.D., and King, C. (2020). The persistence of high energy burdens: A bibliometric analysis of vulnerability, poverty, and exclusion in the united states. Energy Res. Soc. Sci. *70*, 101756. https://doi.org/10.1016/J.ERSS.2020.101756.
20. Cong, S., Nock, D., Qui, Y. L., and Xing, B. (2021). The energy equity gap: Unveiling hidden energy poverty. https://www.researchsquare.comhttps://www.researchsquare.com/article/rs-712945/v1. doi:10.21203/RS.3.RS-712945/V1.
21. Wells, E.M., Berges, M., Metcalf, M., Kinsella, A., Foreman, K., Dearborn, D.G., and Greenberg, S. (2015). Indoor air quality and occupant comfort in homes with deep versus conventional energy efficiency renovations. Build. Environ. *93*, 331–338. https://doi.org/10.1016/J.BUILDENV.2015.06.021.
22. Fabian, P., Adamkiewicz, G., and Levy, J.I. (2012). Simulating indoor concentrations of no2 and pm2.5 in multifamily housing for use in health-based intervention modeling. Indoor Air *22*, 12–23. https://doi.org/10.1111/J.1600-0668.2011.00742.X.
23. Chen, C.-f., Feng, J., Luke, N., Kuo, C.-P., and Fu, J.S. (2022). Localized energy burden, concentrated disadvantage, and the feminization of energy poverty. iScience *25*, 104139. https://doi.org/10.1016/j.isci.2022.104139.
24. Cong, S., Ku, A.L., Nock, D., Ng, C., and Qiu, Y.L. (2024). Comfort or cash? Lessons from the COVID-19 pandemic's impact on energy insecurity and energy limiting behavior in households. Energy Res. Soc. Sci. *113*, 103528. https://doi.org/10.1016/j.erss.2024.103528.
25. Buylova, A. (2020). Spotlight on energy efficiency in oregon: Investigating dynamics between energy use and socio-demographic characteristics in spatial modeling of residential energy consumption. Energy Policy *140*, 111439. https://doi.org/10.1016/J.ENPOL.2020.111439.
26. Bednar, D.J., Reames, T.G., and Keoleian, G.A. (2017). 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. https://doi.org/10.1016/J.ENBUILD.2017.03.028.
27. Moore, D., and Webb, A.L. (2022). Evaluating energy burden at the urban scale: A spatial regression approach in cincinnati, ohio. Energy Policy *160*, 112651. https://doi.org/10.1016/J.ENPOL.2021.112651.
28. Mohai, P., Pellow, D., and Roberts, J.T. (2009). Environmental justice. Annu. Rev. Environ. Resour. *34*, 405–430. https://doi.org/10.1146/annurev-environ-082508-094348.
29. R. Holifield, J. Chakraborty, and G. Walker, eds. (2017). The Routledge Handbook of Environmental Justice, 1st ed. (Routledge). https://doi.org/10.4324/9781315678986.
30. Rentschler, J., and Leonova, N. (2023). Global air pollution exposure and poverty. Nat. Commun. *14*, 4432. https://doi.org/10.1038/s41467-023-39797-4.
31. Hauptman, M., Rogers, M.L., Scarpaci, M., Morin, B., and Vivier, P.M. (2023). Neighborhood disparities and the burden of lead poisoning. Pediatr. Res. *94*, 826–836. https://doi.org/10.1038/s41390-023-02476-7.
32. Hilmers, A., Hilmers, D.C., and Dave, J. (2012). Neighborhood disparities in access to healthy foods and their effects on environmental justice. Am. J. Public Health *102*, 1644–1654. https://doi.org/10.2105/AJPH.2012.300865.
33. Baker, E., Carley, S., Castellanos, S., Nock, D., Bozeman, J.F., Konisky, D., Monyei, C.G., Shah, M., and Sovacool, B. (2023). Metrics for

decision-making in energy justice. Annu. Rev. Environ. Resour. *48*, 737–760. https://doi.org/10.1146/annurev-environ-112621-063400.

34. Do, V., McBrien, H., Flores, N.M., Northrop, A.J., Schlegelmilch, J., Kiang, M.V., and Casey, J.A. (2023). Spatiotemporal distribution of power outages with climate events and social vulnerability in the usa. Nat. Commun. *14*, 2470. https://doi.org/10.1038/s41467-023-38084-6.

35. Macmillan, M., Wilson, K., Baik, S., Carvallo, J.P., Dubey, A., and Holland, C.A. (2023). Shedding light on the economic costs of long-duration power outages: A review of resilience assessment methods and strategies. Energy Res. Soc. Sci. *99*, 103055. https://doi.org/10.1016/j.erss.2023.103055.

36. Chi Hsu, F., Taneji, J., Carvallo, J., and Shah, Z. Frozen out in texas: Blackouts and inequity. https://www.rockefellerfoundation.org/insights/grantee-impact-story/frozen-out-in-texas-blackouts-and-inequity/.

37. Bhattacharyya, A., and Hastak, M. (2023). A data-driven approach to quantify disparities in power outages. Sci. Rep. *13*, 7247. https://doi.org/10.1038/s41598-023-34186-9.

38. Lee, C.-C., Maron, M., and Mostafavi, A. (2022). Community-scale big data reveals disparate impacts of the Texas winter storm of 2021 and its managed power outage. Humanit. Soc. Sci. Commun. *9*, 335. https://doi.org/10.1057/s41599-022-01353-8.

39. Yabe, T., and Ukkusuri, S.V. (2020). Effects of income inequality on evacuation, reentry and segregation after disasters. Trans. Res. D Trans. Environ. *82*, 102260. https://doi.org/10.1016/j.trd.2020.102260.

40. Garland, J., Baker, K., and Livneh, B. (2023). Weather-induced power outage prediction: A comparison of machine learning models. In 2023 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6. https://doi.org/10.1109/SmartGridComm57358.2023.10333953.

41. Baik, S., Davis, A.L., Park, J.W., Sirinterlikci, S., and Morgan, M.G. (2020). Estimating what us residential customers are willing to pay for resilience to large electricity outages of long duration. Nat. Energy *5*, 250–258. https://doi.org/10.1038/s41560-020-0581-1.

42. Coyle, D., and Weller, A. (2020). "explaining" machine learning reveals policy challenges. Science (1979). *368*, 1433–1434. https://doi.org/10.1126/science.aba9647.

43. Innes, J.E. (1990). Knowledge and Public Policy: The Search for Meaningful Indicators (Transaction Publishers).

44. Bell, A., Solano-Kamaiko, I., Nov, O., and Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability tradeoff in machine learning for public policy. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22 (Association for Computing Machinery), pp. 248–266. https://doi.org/10.1145/3531146.3533090.

45. Peet, E.D., Vegetabile, B.G., Cefalu, M., Pane, J.D., Damberg, C.L. (2022). Machine learning in public policy: The perils and the promise of interpretability. https://doi.org/10.7249/PEA828-1.

46. Ghorbany, S., Hu, M., Sisk, M., Yao, S., and Wang, C. (2024). Passive over active: How low-cost strategies influence urban energy equity. Sustain. Cities Soc. *114*, 105723. https://doi.org/10.1016/j.scs.2024.105723.

47. Ghorbany, S., Hu, M., Yao, S., Wang, C., Nguyen, Q.C., Yue, X., Alirezaei, M., Tasdizen, T., and Sisk, M. (2024). Examining the Role of Passive Design Indicators in Energy Burden Reduction: Insights from a Machine Learning and Deep Learning Approach. Build. Environ. *250*, 111126. https://doi.org/10.1016/j.buildenv.2023.111126.

48. Spandagos, C., Tovar Reaños, M.A., and Lynch, M.Á. (2023). Energy poverty prediction and effective targeting for just transitions with machine learning. Energy Econ. *128*, 107131. https://doi.org/10.1016/j.eneco.2023.107131.

49. Wang, Y., Chen, X., Gao, M., and Dong, J. (2022). The use of random forest to identify climate and human interference on vegetation coverage changes in southwest china. Ecol. Indic. *144*, 109463. https://doi.org/10.1016/j.ecolind.2022.109463.

50. Simon, S.M., Glaum, P., and Valdovinos, F.S. (2023). Interpreting random forest analysis of ecological models to move from prediction to explanation. Sci. Rep. *13*, 3881. https://doi.org/10.1038/s41598-023-30313-8.

51. Mallala, B., Ahmed, A.I.U., Pamidi, S.V., Faruque, M.O., and M, R.R. (2025). Forecasting global sustainable energy from renewable sources using random forest algorithm. Results Eng. *25*, 103789. https://doi.org/10.1016/j.rineng.2024.103789.

52. Bienvenido-Huertas, D., Pulido-Arcas, J.A., Rubio-Bellido, C., and Pérez-Fargallo, A. (2021). Prediction of fuel poverty potential risk index using six regression algorithms: A case-study of chilean social dwellings. Sustainability *13*, 2426. https://doi.org/10.3390/su13052426.

53. Hong, Z., and Park, I.K. (2021). Comparative analysis of energy poverty prediction models using machine learning algorithms. jkpa. *56*, 239–255. https://doi.org/10.17208/jkpa.2021.10.56.5.239.

54. Balkissoon, S., Fox, N., Lupo, A., Haupt, S.E., Penny, S.G., Miller, S.J., Beetstra, M., Sykuta, M., and Ohler, A. (2024). Forecasting energy poverty using different machine learning techniques for missouri. Energy *313*, 133904. https://doi.org/10.1016/j.energy.2024.133904.

55. Wang, H., Maruejols, L., and Yu, X. (2021). Predicting energy poverty with combinations of remote-sensing and socioeconomic survey data in india: Evidence from machine learning. Energy Econ. *102*, 105510. https://doi.org/10.1016/j.eneco.2021.105510.

56. United States Census Bureau, C. R. E. About community resilience estimates. Available at https://www.census.gov/programs-surveys/community-resilience-estimates/about.html(2021/11/26).

57. United States Environmental Protection Agency, E. Ejscreen: Environmental justice screening and mapping tool. Available at https://www.epa.gov/ejscreen(2021/11/26).

58. Oceanic, C.-N., and Atmospheric Administration, N. https://www.noaa.gov/climate.

59. Wilson, E.J.H., Harris, C.B., Robertson, J.J., and Agan, J. (2019). Evaluating energy efficiency potential in low-income households: A flexible and granular approach. Energy Policy *129*, 710–737. https://doi.org/10.1016/j.enpol.2019.01.054.

60. Brelsford, C., Tennille, S., Myers, A., Chinthavali, S., Tansakul, V., Denman, M., Coletti, M., Grant, J., Lee, S., Allen, K., et al. (2024). A dataset of recorded electricity outages by united states county 2014–2022. Sci. Data *11*, 271. https://doi.org/10.1038/s41597-024-03095-5.

61. van Hove, W., Dalla Longa, F., and van der Zwaan, B. (2022). Identifying predictors for energy poverty in europe using machine learning. Energy Build. *264*, 112064. https://doi.org/10.1016/j.enbuild.2022.112064.

62. Grzybowska, U., Wojewódzka-Wiewiórska, A., Vaznonienė, G., and Dudek, H. (2024). Households vulnerable to energy poverty in the visegrad group countries: An analysis of socio-economic factors using a machine learning approach. Energies *17*, 6310. https://doi.org/10.3390/en17246310.

63. Satapathy, S.K., Saravanan, S., Mishra, S., and Mohanty, S.N. (2023). A comparative analysis of multidimensional covid-19 poverty determinants: An observational machine learning approach. New Gener. Comput. *41*, 155–184. https://doi.org/10.1007/s00354-023-00203-8.

64. Ortiz, L., Gamarro, H., Gonzalez, J.E., and McPhearson, T. (2022). Energy burden and air conditioning adoption in new york city under a warming climate. Sustain. Cities Soc. *76*, 103465. https://doi.org/10.1016/J.SCS.2021.103465.

65. Subakti, D. Managing the july 2024 heat wave with our partners in california and the west (2024). https://www.caiso.com/about/news/managing-the-july-2024-heat-wave-with-our-partners-in-california-and-the-west.

66. Wang, H., and Chen, Q. (2014). Impact of climate change heating and cooling energy use in buildings in the united states. Energy Build. *82*, 428–436. https://doi.org/10.1016/J.ENBUILD.2014.07.034.

67. Ma, O., Krystal Laymon, R. O. J. W., M. Day, and Vimont, A. Low-income energy affordability data (lead) tool methodology. Available at https://lead.openei.org/assets/docs/LEAD-Tool-Methodology.pdf(2021/12/08).
68. Abolafia-Rosenzweig, R., He, C., and Chen, F. (2022). Winter and spring climate explains a large portion of interannual variability and trend in western u.s. summer fire burned area. Environ. Res. Lett. *17*, 054030. https://doi.org/10.1088/1748-9326/ac6886.
69. Berg, K., Kuhn, S., and Van Dyke, M. (2017). Spatial surveillance of childhood lead exposure in a targeted screening state: An application of generalized additive models in denver, colorado. J. Public Health Manag. Pract. *23*, S79–S92. https://doi.org/10.1097/PHH.0000000000000620.
70. Kessels, R., Hoornweg, A., Thanh Bui, T.K., and Erreygers, G. (2020). A distributional regression approach to income-related inequality of health in Australia. Int. J. Equity Health *19*, 102. https://doi.org/10.1186/s12939-020-01189-1.
71. Sundararajan, A., and Ollis, B. (2021). Regression and generalized additive model to enhance the performance of photovoltaic power ensemble predictors. IEEE Access *9*, 111899–111914. https://doi.org/10.1109/ACCESS.2021.3103126.
72. Ravindra, K., Rattan, P., Mor, S., and Aggarwal, A.N. (2019). Generalized additive models: Building evidence of air pollution, climate change and human health. Environ. Int. *132*, 104987. https://doi.org/10.1016/j.envint.2019.104987.
73. Kocaguneli, E., and Menzies, T. (2013). Software effort models should be assessed via leave-one-out validation. J. Syst. Software *86*, 1879–1890. https://doi.org/10.1016/j.jss.2013.02.053.
74. Scheier, E., and Kittner, N. (2022). A measurement strategy to address disparities across household energy burdens. Nat. Commun. *13*, 288. https://doi.org/10.1038/s41467-021-27673-y.
75. Dugan, J., Byles, D., and Mohagheghi, S. (2023). Social vulnerability to long-duration power outages. Int. J. Disaster Risk Reduct. *85*, 103501. https://doi.org/10.1016/j.ijdrr.2022.103501.
76. Jeffers, R.F., Baca, M.J., Wachtel, A.M., DeRosa, S., Staid, A., Fogleman, W.E., Outkin, A.V., and Currie, F.M.. Analysis of Microgrid Locations Benefitting Community Resilience for Puerto Rico. https://doi.org/10.2172/1481633.
77. Casey, J.A., Mango, M., Mullendore, S., Kiang, M.V., Hernández, D., Li, B.H., Li, K., Im, T.M., and Tartof, S.Y. (2021). Trends from 2008 to 2018 in electricity-dependent durable medical equipment rentals and sociodemographic disparities. Epidemiology *32*, 327–335. https://doi.org/10.1097/EDE.0000000000001333.
78. Brown Weiss, E. (2008). Climate Change, Intergenerational Equity, and International Law (Georgetown Law Faculty Publications and Other Works). https://scholarship.law.georgetown.edu/facpub/1625.
79. Agency, U. E. P. Climate change and the health of socially vulnerable people. https://www.epa.gov/climateimpacts/climate-change-and-health-socially-vulnerable-people#8foot.
80. Ashbaugh, M., and Kittner, N. (2024). Addressing extreme urban heat and energy vulnerability of renters in portland, or with resilient household energy policies. Energy Policy *190*, 114143. https://doi.org/10.1016/j.enpol.2024.114143.
81. Swanwick, R.H., Read, Q.D., Guinn, S.M., Williamson, M.A., Hondula, K. L., and Elmore, A.J. (2022). Dasymetric population mapping based on us census data and 30-m gridded estimates of impervious surface. Sci. Data *9*, 523. https://doi.org/10.1038/s41597-022-01603-z.
82. Bureau, U. C. (2022). geographic levels. https://www.census.gov/programs-surveys/economic-census/geographies/levels/2022-levels.html.
83. Meng, Y., Cave, M., and Zhang, C. (2019). Comparison of methods for addressing the point-to-area data transformation to make data suitable for environmental, health and socio-economic studies. Sci. Total Environ. *689*, 797–807. https://doi.org/10.1016/j.scitotenv.2019.06.452.
84. Resstock geographic fields and codes. https://github.com/NREL/ResStock.github.io/ResStock.github.io/docs/resources/explanations/Geographic_Fields_and_Codes.html.
85. Bureau, U. C. Geographic areas reference manual. https://www.census.gov/programs-surveys/geography/guidance/geographic-areas-reference-manual.html.
86. Mastropietro, P., Rodilla, P., and Batlle, C. (2020). Emergency measures to protect energy consumers during the covid-19 pandemic: A global review and critical analysis. Energy Res. Soc. Sci. *68*, 101678. https://doi.org/10.1016/J.ERSS.2020.101678.
87. Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography *36*, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.
88. Lumley, T., and Miller, A. leaps: Regression subset selection (1997). https://cran.r-project.org/web/packages/leaps/leaps.pdf. https://doi.org/10.32614/CRAN.package.leaps.
89. Patrizio, P., Pratama, Y.W., and Dowell, N.M. (2020). Socially equitable energy system transitions. Joule *4*, 1700–1713. https://doi.org/10.1016/j.joule.2020.07.010.
90. Morrison, D.R., Jacobson, S.H., Sauppe, J.J., and Sewell, E.C. (2016). Branch-and-bound algorithms: A survey of recent advances in searching. branching, and pruning *19*, 79–102. https://doi.org/10.1016/j.disopt.2016.01.005.
91. Wood, S. (2017). Generalized Additive Models: An Introduction with R, Second Edition (2nd ed. (Chapman and Hall/CRC). https://doi.org/10.1201/9781315370279.
92. Wood, S. R: Generalized additive model selection. https://astrostatistics.psu.edu/su07/R/library/mgcv/html/gam.selection.html.
93. Banerjee, M., Ding, Y., and Noone, A.-M. (2012). Identifying representative trees from ensembles. Stat. Med. *31*, 1601–1616. https://doi.org/10.1002/sim.4492.
94. Representative trees from ensembles (2024). https://github.com/araastat/reprtree.
95. Jiang, Y., Li, Z., and Cutter, S.L. (2021). Social distance integrated gravity model for evacuation destination choice. Int. J. Dig. Earth *14*, 1004–1018. https://doi.org/10.1080/17538947.2021.1915396.
96. Pandas: Python data analysis library. https://pandas.pydata.org/.
97. Tidyverse packages. https://www.tidyverse.org/packages/.
98. Create elegant data visualisations using the grammar of graphics. https://ggplot2.tidyverse.org/.
99. Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2002). randomforest: Breiman and Cutlers Random Forests for Classification and Regression. https://CRAN.R-project.org/package=randomForest.doi:10.32614/CRAN.package.randomForestinstitution: Comprehensive R Archive Network.
100. Wood, S. (2000). mgcv: Mixed Gam Computation Vehicle with Automatic Smoothness Estimation. https://CRAN.R-project.org/package=mgcv.doi:10.32614/CRAN.package.mgcvinstitution: Comprehensive R Archive Network.
101. based on Fortran code by Alan Miller, T. L. leaps: Regression Subset Selection (1997). https://CRAN.R-project.org/package=leaps.doi:10.32614/CRAN.package.leapsinstitution: Comprehensive R Archive Network.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Energy Burden	Low-Income Energy Affordability Data (LEAD), The United States Department of Energy: https://lead.openei.org/assets/docs/LEAD-Tool-Methodology.pdf	
Count of housing units built before 1960, Diesel particulate matter level in air, Air toxics cancer risk, Air toxics respiratory hazard index, Traffic proximity and volume, Indicator for major direct dischargers to water, Proximity to national priorities list (NPL) sites, Proximity to risk management plan (RMP) facilities, Proximity to treatment storage and disposal (TSDF) facilities, Ozone level in air, PM2.5 level in air.	Environmental Justice Screening Tool (EJ Screen), The United States Environmental Protection Agency): https://www.epa.gov/ejscreen	
Estimated number of individuals with zero risk factors, Estimated number of individuals with one-two risk factors, Estimated number of individuals with three plus risk factors.	Community Resilience for Equity and Disasters, United States: https://www.census.gov/programs-surveys/community-resilience-estimates/data/datasets.html	
A/C type in home, Type of Home, Number of bedrooms in home, Energy Source	ResStock, The National Renewable Energy Laboratory: https://resstock.nrel.gov/datasets	
Power Outages	A dataset of recorded electricity outages by United States county 2014–2022, https://doi.org/10.1038/s41597-024-03095-5	
Data, code, and other items generated items	Available by request from lead author	
Software and algorithms		
R Programming	https://www.r-project.org/about.html	
Python Programming	https://www.python.org	

METHOD DETAILS

This section provides an overview of the methodology implemented in this paper. First, the data acquisition and processing are discussed, which explains the creation of the data values and the PC datasets. Figure 2 shows that the modeling frameworks use data values and PCs. Following the data processing, the modeling frameworks are discussed, followed by the evaluation criteria, which provide additional metrics to support the main body of the manuscript.

Data acquisition and processing

This study considers the contiguous U.S. by census region. Census regions are four groups of states established by the Census Bureau in 1942. The four groups are the Northeast, South, Midwest, and West. Census regions provide geographic frameworks at larger scales to perform statistical analysis, such as the work completed in this study, summarize data, and offer varying physical and cultural geography.[85] The data used in this study is at the county level and can be categorized as weather and location, building characteristics, environmental justice (demographic and environmental), community resilience, power outages, and energy burden. The weather and geographic data was acquired from the NOAA National Centers for Environmental Information Climate online data tool. For this study, a mean temperature for the month of July was used to represent a summer month. To represent the U.S.

residential building stock, building characteristics were obtained from the NREL tool ResStock. ResStock is a residential building stock model that simulates the diversity of the residential housing stock in the U.S.[59] Housing stock metrics used in this study include the type of A/C in the home, since weather from a summer month is used, energy expenditures per household, and the building type.

Since the building characteristic data is categorical, this data was one hot encoded and then summed for each category. For instance, there are four categories relating to A/C type. Each occurrence of an A/C type is summed and multiplied by the weighting factor in ResStock to be representative of the households in each respective county. Due to the COVID-19 pandemic resulting in many individuals partaking in increased home activities, working, and schooling from home, this often means higher energy bills. Given these continued stresses due to the COVID-19 pandemic, researchers have found evidence linking high energy burdens with conditions that may increase a household's vulnerability to COVID-19 and related psychological stresses due to potential evictions or loss of electricity due to defaulting payments.[86]

To account for these stresses, the community resilience for equity and disasters tool, developed by the U.S. Census Bureau in 2020, is used. The tool was inspired by how COVID-19 was disproportionately impacting minority communities. Thus, the intent of this tool is to measure the capacity of individuals and households to recover from stresses such as local health or environmental disasters.[56] The EPA Environmental Justice screening tool (EJ Screen) was used to acquire environmental justice data. EJ screen features 11 environmental indicators and six demographic indicators.[57] The environmental indicators can be further broken down into three sub-groups: Potential Exposure (Lead paint, Ozone, etc), Proximity (traffic and volume, NPL sites, etc.), and Hazard/Risk (air toxic cancer and respiratory risk). The power outage data was taken from Brelsford et al.,[60] and a five year average from 2016 through 2020 was used, to match the energy burden data. Data for energy burden used in this study is from the DOE Low-Income Energy Affordability Data. This data was created to increase awareness of low-income household issues relating to energy. For spatial allocations of different housing units, an iterative proportional fitting algorithm was used with survey-based residential energy consumption cross-tabulations from the U.S. Census housing data from the 2016 five-year American Community Survey.[67] An overview of the data is provided in Table 5.

Table 5. Indicator descriptions per data source

Data Source	Data description
Low-income energy affordability data (LEAD) (Ma et al.[67])	The average county energy burden (%). The average was taken from the years 2016–2020.
Environmental justice (United States Environmental Protection Agency[57])	The EPA's EJ Screen data represents the environmental justice parameters including 11 environmental indicators and six demographic indicators \citep{ej_data}. The environmental indicators can be further broken down into three sub-groups: potential exposure (lead paint, ozone, etc), proximity (traffic and volume, NPL sites, etc.), and hazard/risk (air toxic cancer and respiratory risk).
Community resilience for equity and disasters (United States Census Bureau[56])	From the U.S. Census Bureau, these data represents the estimated number of individuals per county that experience either zero risk (no risk/low risk), one-two risk (low risk/moderate risk), or three risk (high risk).
Building characteristics (Brelsford et al.[60])	The NREL tool ResStock is used to represent the residential building characteristics. Housing stock metrics used in this study include the type of cooling in the home, since weather from a summer month is used, the year the dwelling was built, energy expenditures per household, and the building type.
Location and temperature (Oceanic and Atmospheric Administration[58])	From NOAA, a mean temperature for the month of July was used to represent a summer month outdoor temperature.
Power outages (Brelsford et al.[60])	The average power outage duration, number of customers impacted, and occurrence for the years 2016–2020.

A brief conceptual description of each data source is described.

Principle component analysis

The data used had a total of 42 indicators that could be used to predict or describe energy burden. As such, dimensionality reduction in the form of a PCA was completed using singular value decomposition. PCA is a common method used to remove correlations and reduce collinearity, as correlation may indicate collinearity.[87] As such, PCA outputs an orthogonal axes, allowing for the PCs to be directly used in the model in place of the original data. Using PCs in the model additionally increases computational efficiency while allowing information from more indicators to be considered in the models, without increasing model configurations.

Data modeling

This section provides an overview of the data modeling techniques used in this paper. A mathematical representation may be found in the supplemental information Methods S2. The stepwise subset selection, GAMs, random forests, LOCO are performed on the full indicator set and the PCs.

Stepwise subset selection

To reduce the number of indicators from 42 to 15 for each region, a ESS, in the form of an exhaustive search using a "Branch and Bound" algorithm, was performed.[88] Datasets using every combination of five indicators from the 15 indicators selected by the ESS are then created (3,003 datasets). A reduction in the number of indicators used in the modeling frameworks was performed as the question, "How to appropriately create equitable and just energy policies?" is largely unknown.[89] Thus, a smaller subset of data was used to increase model interpretability and to show the importance of individual attributes that are selected as the most important indicators in a more detailed manner for policy decisions. To, in return, foster a deeper understanding of the energy burden within the general public and governing bodies. In the supplemental information Methods S2, pseudo-code adapted from[90] is provided to show an example of the "Branch and Bound" algorithm. Conceptually, the search space (combinations of the 42 data variables) is optimized (find the combination of 15 that results in the lowest akaike information criterion score) by intelligently testing all possible solutions. Intelligent testing is the ability to prune, as when a lower bound of a branch does not perform as well as the current best solution, that branch (or subset) is automatically disregarded since it will never produce the optimal solution. To expand upon the results in the main body of this work, Table 6 shows the variables selected and used throughout this study, while Table 7 shows the variables that would have been selected if income related variables are removed.

Table 6. Exhaustive search subset selection methods

Midwest	Northeast	South	West
Population	Poverty status	Low-income individuals	Poverty status
Low-income individuals	Low-income individuals	Traffic proximity and volume	Low-income individuals
Traffic proximity and volume	Median age	Poverty status	Median age
Median age	Air toxics respiratory hazard	Ozone level in air	Air toxics respiratory hazard
Proximity to risk management plan facilities	Natural gas consumption	Minority population	Natural gas consumption
1-2 Bedroom home	Propane consumption	Median age	Propane consumption
Households in linguistic isolation	Traffic proximity and volume	Households in linguistic isolation	Diesel particulate matter
Natural gas consumption	Diesel particulate matter	Population	Traffic proximity and volume
Fuel oil consumpation	Ozone level in air	Mobile home	Community resilience: low risk
Power outage duration	Community resilience: low risk	Proximity to risk management plan facilities	Ozone level in air
Less than high school degree	Less than high school degree	Propane consumption	1-2 Bedroom home
Proximity to hazardous waste	1-2 Bedroom home	Air toxics respiratory hazard	Less than high school degree
Minority population	Multifamily housing	Less than high school degree	Multifamily housing
Air toxics respiratory hazard	Average power outage occurance	Homes built prior to 1960	Households in linguistic isolation
PM 2.5 concentration	Households in linguistic isolation	Proximity to national priorities list sites	Homes built prior to 1960

The indicators selected for each region from the exhaustive search subset selection using all data values.

Table 7. Exhaustive search subset selection results alternative

Midwest	Northeast	South	West
Population	*Minority population	*Multifamily housing	*Population
Less than high school degree	*Population	Traffic proximity and volume	*Air toxins related to cancer risk
Traffic proximity and volume	Median age	*Natural gas consumption	Median age
Median age	*Homes built prior to 1960	Ozone level in air	*Outside Temperature
Proximity to risk management plan facilities	*Air toxins related to cancer risk	Minority population	*Having room A/C in home
1-2 Bedroom home	Propane consumption	Median age	Propane consumption

(*Continued on next page*)

Table 7. *Continued*

Midwest	Northeast	South	West
Households in linguistic isolation	Traffic proximity and volume	Households in linguistic isolation	*PM 2.5 in air
Natural gas consumption	Ozone level in air	Population	Traffic proximity and volume
Fuel oil consumpation	*PM 2.5 in air	*Residential heat pump	*Major direct dischargers to water
Power outage duration	*Having room A/C in home	Proximity to risk management plan facilities	Ozone level in air
Less than high school degree	*Major direct dischargers to water	Propane consumption	1-2 Bedroom home
*Multifamily housing	1-2 Bedroom home	Air toxics respiratory hazard	*Minority Population
Minority population	Multifamily housing	Less than high school degree	Multifamily housing
*Community resilience: low risk	Average power outage occurrence	Homes built prior to 1960	*Proximity to risk management plan facilities
*Diesel particulate matter in air	*Proximity to risk management plan facilities	Proximity to national priorities list sites	Homes built prior to 1960

The indicators selected for each region from the exhaustive search subset selection when low-income individuals and poverty status are not considered.

Generalized additive model

GAMs are data-driven rather than model-driven methods due to the data determining the relationship between indicators and response variables. Thus, no parametric relationship is assumed. Instead, GAMs are semi-parametric extensions of generalized linear models (GLMs). This relationship is shown in the supplemental information Methods S2. To summarize, GAMs substitute the linear terms $\sum \beta_j x_j$ with an additive (summation) of nonlinear smooth functions $\sum S_j(x_j)$.

An advantage of GAMs is their ability to handle nonlinear and non-monotonic relationships between the indicators and the response variable, which assists the model in better representing the data, and have been found to outperform more complex black box models, such as random forest.[68] GAMs have been used in various studies and fields, including studying lead exposure in children,[69] income inequality in healthcare,[70] renewable energy power production,[71] and climate relations to air quality and health.[72] Thus, GAMs hold eminent potential in predicting and understanding the complex interactions between the data groups and energy burden.

Specific to this study, 20 equations test different combinations of smoothing spline functions and polynomials. These equations are provided in Methods S2 and Table S1. Thus, testing each of the 3,003 datasets for each GAM creates 60,006 models per region. The top 100 performing models are selected based on the GCV Un-Biased Risk Estimator (UBRE) score. The UBRE can be thought of as a scaled AIC score or Mallows C_p for an additive model.[91] A smaller UBRE indicates a better model fit when considering precision and bias. When comparing GAMs, the UBRE is recommended and is the most consistently used method for GAM model comparison.[91,92] After the top-performing 100 models are selected, a LOCO analysis is completed.

Random forest

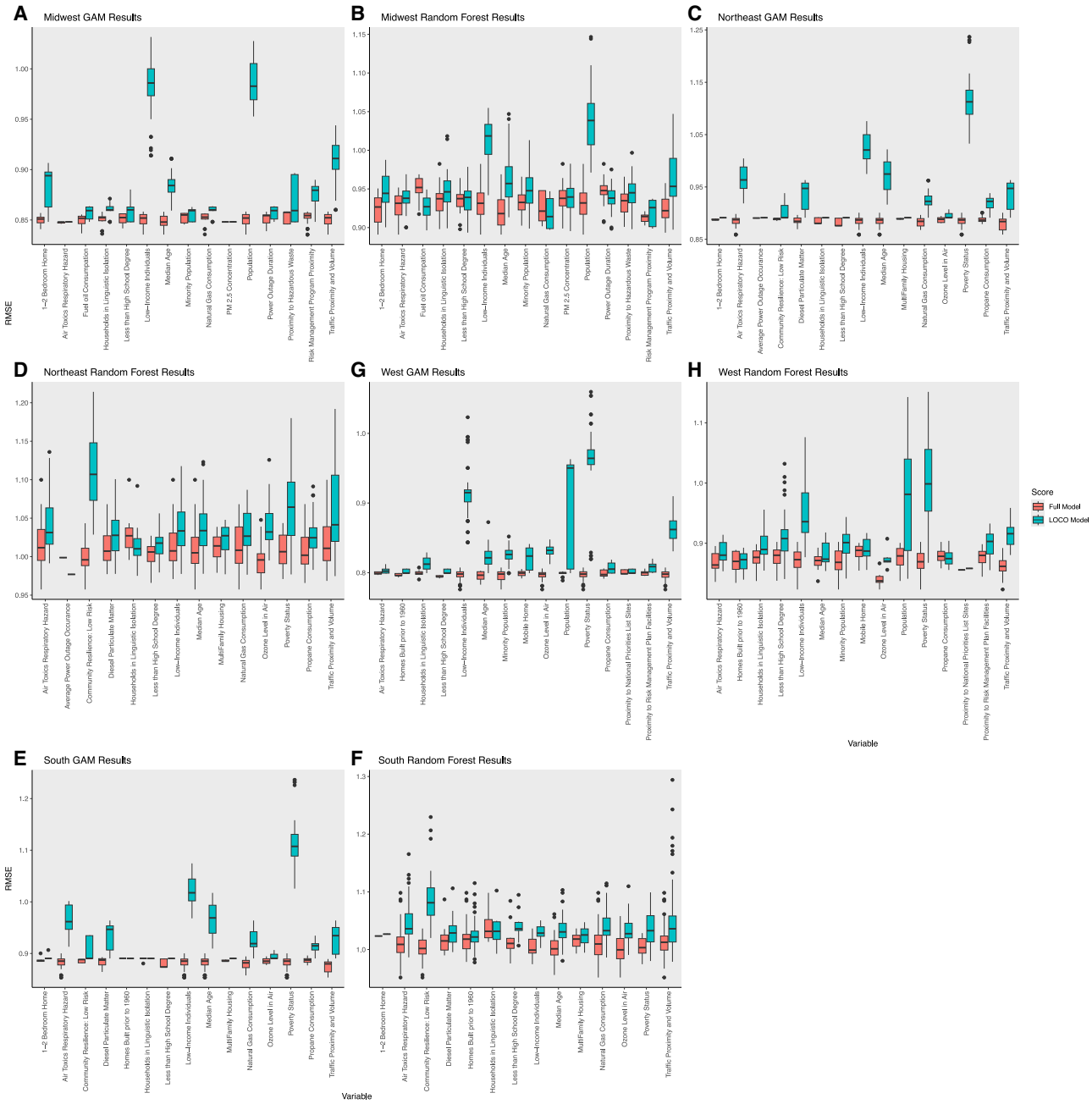
Random forest is an extension of bagging or bootstrap aggregating methods, meaning that a number of repeated samples, B, of the training data are used. The "Forest" is created as an ensemble of decision trees. Each decision tree is trained on a subset, *bth* set of different indicators, to create diversity and aid in robustness in the "Forest". The final step is the model prediction, which uses an average among the individual decision trees to create a single model with low variance. In this study, 500 decision trees are created within the random forest. To determine the best-performing random forest models the R^2 was used. Note that this does differ for a classification problem. Random forests offer greater predictive power than decision trees, although decision trees are more intuitive since they are not a black box method like random forests. Since model interpretability is lost in random forest models, a single tree that is representative of the forest is created using methods from,[93,94] where a d_2 metric that represents the closeness based on prediction is used, for regression this is the euclidean distance.

Leave-one-column-out

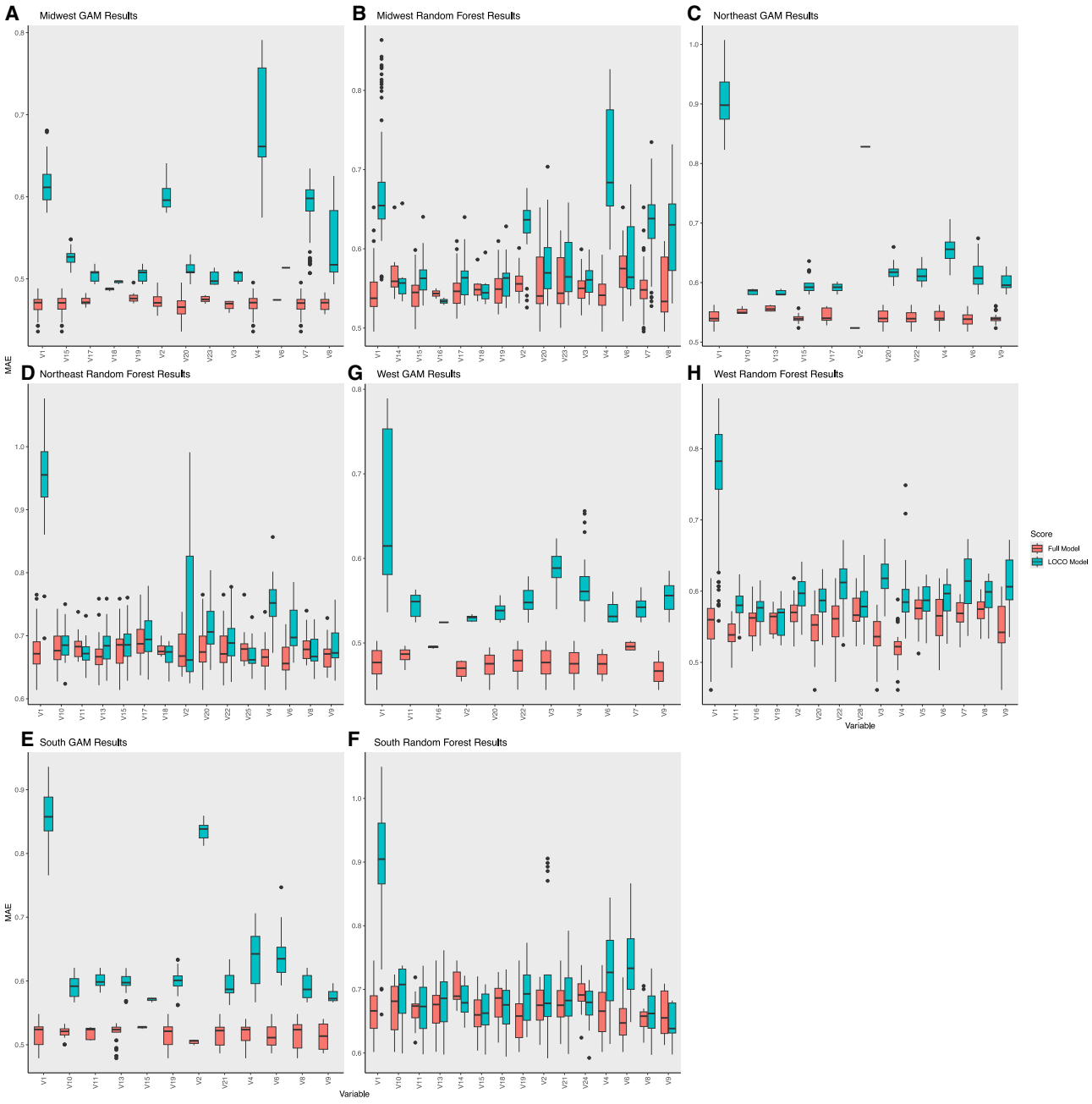
The LOCO analysis considers the indicator influence of the top-performing models from frameworks one and two. It is iterative, as each of the top-performing models is re-evaluated with one indicator left out. The model with the indicator left out is then compared to the model with all indicators included. LOCO analysis are completed to understand the individual indicators' effect on the overall estimate or prediction.[68,95]

Evaluation criteria

To assess the fit of the models, both modeling frameworks used the R^2, which is found in the main body of the paper in Figures 4, 5, 6, and 7, RMSE, and MAE. The R^2 is used to understand the proportion of energy burden variance explained by the indicators. The RMSE is one the most common metrics used for evaluating the predictive qualities of a model, and measures the distance from the predicted value to the actual value. The RMSE is provided in Figure 8. The MAE measure of the average size of the mistakes in a collection of predictions without considering the direction (positive or negative). The is provided in Figure 8.



Leave-one-column-out root-mean-square error



LOCO mean absolute error for the PCs

QUANTIFICATION AND STATISTICAL ANALYSIS

This study did not use human, animal, or plant subjects or conduct statistical experiments on human, animal, or plant data. The data cleaning, was completed in Python, using the PANDAS package.[96] The models, analysis, and plotting were completed using R programming language including tidyverse,[97] ggplot,[98] randomForest,[99] mgcv,[100] and leaps.[101]