Machine Learning from Schools about Energy Efficiency

Fiona Burlig
University of Chicago

Christopher Knittel
MIT

David Rapson
UC Davis

Mar Reguant
Northwestern University

Catherine Wolfram*
UC Berkeley

April 17, 2019

Abstract

We use high-frequency panel data on electricity consumption to study the effectiveness of energy efficiency upgrades in K-12 schools in California. Using a panel fixed effects approach, we find that these upgrades deliver between 11 and 90 percent of expected savings, depending on specification and treatment of outliers. Using machine learning to inform our specification choice, we estimate a narrower range: 65 to 102 percent, with a central estimate of 78 percent. These results imply that upgrades are performing less well than *ex ante* predictions on average, although we can reject the very low realization rates found in prior work.

JEL Codes: Q4, Q5, C4

Keywords: energy efficiency; machine learning; schools

*Burlig: Harris School of Public Policy and Energy Policy Institute, University of Chicago, burlig@uchicago.edu. Knittel: Sloan School of Management and Center for Energy and Environmental Policy Research, MIT and NBER, knittel@mit.edu. Rapson: Department of Economics, UC Davis, dsrapson@ucdavis.edu. Reguant: Department of Economics, Northwestern University, CEPR and NBER, mar.reguant@northwestern.edu. Wolfram: Haas School of Business and Energy Institute at Haas, UC Berkeley and NBER, cwolfram@berkeley.edu. We thank Dan Buch, Arik Levinson, and Ignacia Mercadal, as well as seminar participants at the Energy Institute at Haas Energy Camp, MIT, Harvard, the Colorado School of Mines, the University of Arizona, Arizona State University, Texas A & M, Iowa State, Boston College, the University of Maryland, Kansas State, Yale University, Columbia University, University of Warwick, the University of Virginia, New York University, the University of Pennsylvania, Carnegie Mellon, the 2016 NBER Summer Institute, and the Barcelona GSE Summer Forum for useful comments. We thank Joshua Blonz and Kat Redoglio for excellent research assistance. We gratefully acknowledge financial support from the California Public Utilities Commission. Burlig was generously supported by the National Science Foundation's Graduate Research Fellowship Program under Grant DGE-1106400. All remaining errors are our own.

1 Introduction

Energy efficiency is a cornerstone of global greenhouse gas (GHG) abatement efforts. For example, worldwide proposed climate mitigation plans rely on energy efficiency to deliver 42 percent of emissions reductions (International Energy Agency (2015)). The appeal of energy efficiency investments is straightforward: they may pay for themselves by lowering future energy bills. At the same time, lower energy consumption reduces reliance on fossil fuel energy sources, providing the desired GHG reductions. A number of public policies—including efficiency standards, utility-sponsored rebate programs, and information provision requirements—aim to encourage more investment in energy efficiency.

Policymakers are likely drawn to energy efficiency because a number of analyses point to substantial unexploited opportunities for cost-effective investments (see, e.g., McKinsey & Company (2009)). Indeed, it is not uncommon for analyses to project that the lifetime costs of these investments are negative. One strand of the economics literature has attempted to explain why consumers might fail to avail themselves of profitable investment opportunities (see, e.g., Allcott and Greenstone (2012), Gillingham and Palmer (2014), and Gerarden, Newell, and Stavins (2015)). The most popular explanations have emphasized the possibility of market failures, such as imperfect information, capital market failures, split incentive problems, and behavioral biases, including myopia, inattentiveness, prospect theory, and reference-point phenomena.

A second strand of literature seeks to better understand the real-world savings and costs of energy efficiency investments. Analyses such as McKinsey & Company (2009) are based on engineering estimates of both the investment costs and the potential energy savings over time rather than field evidence. There are a variety of reasons why these engineering estimates might understate the costs consumers face or overstate savings.[1] Economists have pointed out that accurately measuring the savings from energy efficiency investments is difficult as it requires constructing a counterfactual energy consumption path from which reductions caused by the efficiency investments can be measured (Joskow and Marron (1992)). Recent studies use both experimental (e.g., Fowlie, Greenstone, and Wolfram (2018), Allcott and Greenstone (2017)) and quasi-experimental (e.g., Levinson (2016a), Myers (2015), and Davis, Fuchs, and Gertler (2014)) approaches to developing this counterfactual. These studies, all of which estimate the effectiveness of energy efficiency upgrades in residential settings, find substantial underperformance, with upgrades delivering between 25 and 58 percent of *ex ante* expected savings.

A more complete view of which energy efficiency opportunities are cost-effective requires more

1. For example, engineering models do not take consumer behavior into account. If an energy efficiency upgrade lowers the effective price of energy services and consumers respond by demanding more energy services, the energy efficiency upgrade will look less effective than the engineering prediction – even if this prediction would have been correct in the absence of behavior change. Furthermore, *ex post* evaluation is relatively uncommon in the energy efficiency industry, so there is limited feedback between real-world outcomes – which capture consumer actions – and engineering models (Fowlie, Greenstone, and Wolfram (2018)).

evidence from a variety of settings. While 37 percent of electricity use in the United States in 2014 was residential, over half is attributable to commercial and industrial uses (Energy Information Administration (2015)). Despite the large role of non-household sectors in energy use, however, the existing literature is largely focused on residential energy efficiency (Kushler (2015)).[2] We extend this work into a non-residential sector by estimating the impacts of energy efficiency upgrades in K-12 schools in California from 2008 to 2014. We match electricity consumption data from public K-12 schools in California to energy efficiency upgrade records, and exploit temporal and cross-sectional variation to estimate the causal effect of the energy efficiency investments on energy use, leveraging high-frequency electricity consumption data generated from advanced metering infrastructure ("smart metering").[3]

We estimate two empirical models. The first is a panel data model that uses a rich set of fixed effects and controls to non-parametrically separate the causal effect of energy efficiency upgrades from other confounding factors. We find evidence that our panel fixed effects approach is sensitive to outliers and to specification. However, choosing the "correct" set of controls is difficult as a result of the richness of our data: there are millions of possible candidate covariates, especially once we allow for interactions between control variables and unit or time fixed effects.

To overcome these challenges, we estimate a second empirical model based on new techniques in machine learning. Machine learning methods are increasingly popular in economics and other social sciences. They have been used to predict poverty and wealth (Blumenstock, Cadamuro, and On (2015), Engstrom, Hersh, and Newhouse (2016), Jean et al. (2016)), improve municipal efficiency (Glaeser et al. (2016)), understand perceptions about urban safety (Naik, Raskar, and Hidalgo (2015)), improve judicial decisions to reduce crime (Kleinberg et al. (2017)), and more. We combine our high-frequency electricity consumption data with machine learning methods in order to select among the set of possible covariates in a disciplined manner.

In our machine learning approach, we use each individual school's pre-treatment data only to build a machine learning model of that school's energy consumption. We use LASSO, as well as a set of alternative algorithms, to build these prediction models while avoiding overfitting. We then use each school's model to forecast counterfactual energy consumption in the post-treatment period. These models provide us with a prediction of what would have happened in the absence of any energy efficiency investments in a flexible, data-driven way, allowing us to control parsimoniously for school-specific heterogeneity while enabling systematic model selection. In order to account for common shocks, we then embed these school-by-school counterfactuals in a panel fixed effects model to estimate causal effects.

The identifying assumption for the standard panel fixed effects model and our machine learning

2. A notable exception is Ryan (2018), which studies energy audits in Indian manufacturing firms, and finds evidence of substantial rebound: treated firms use 9.5 percent more electricity.

3. Over 50 percent of US households had smart meters as of 2019. Smartmeter deployments are predicted to increase by over a third by 2020 (Cooper (2016)).

augmented version is the same: that, conditional on a chosen set of controls, treated schools would have continued on a parallel trajectory to untreated schools in the absence of treatment. We provide evidence in support of these assumptions by demonstrating that treated and untreated schools do not exhibit differential trends in school characteristics, and by showing that there is a trend break among treated schools at the time of treatment. In this selection-on-unobservables design, our machine learning framework allows us to select a richer set of control variables in a systematic and computationally tractable manner.[4]

Using our machine learning method, we find that energy efficiency investments installed in California's K-12 schools underperform relative to average *ex ante* engineering projections of expected savings. Using our machine learning approach, we find that the average energy upgrade delivers approximately 78 percent of expected savings. Importantly, we cannot always reject the hypothesis that these investments deliver the full expected benefits. Comparing our machine learning approach to standard panel fixed effects approaches yields two primary findings. First, we show that estimates from standard panel fixed effects approaches are quite sensitive to specification, outliers, and the set of untreated schools we include in our models, with estimated energy savings ranging from 11 to 90 percent of *ex ante* expectations. Second, by contrast, our machine learning method yields estimates that are substantially more stable across specifications and samples: we estimate savings between 65 and 102 percent of *ex ante* expectations. In addition to enabling data-driven covariate choice, these results highlight another potential benefit of using machine learning.

We explore the extent to which we are able to predict realization rates using easily-observable characteristics. We find suggestive evidence that heating, ventilation, and air conditioning (HVAC) and lighting interventions, which together make up 74 percent of upgrades, are more effective. We also find that larger schools achieve higher realization rates. Though these estimates are noisy and we cannot rule out these schools are simply different from their smaller counterparts, policymakers may be able to make progress towards identifying schools where upgrades are more effective. Finally, although we are substantially limited by our data to perform a full cost-benefit analysis, we discuss the implications of our estimated realization rates in terms of policy evaluation.

The remainder of this paper proceeds by describing our empirical setting and data (Section 2). We then describe the baseline panel fixed approach methodology and present realization rate estimates using these standard tools (Section 3.1). Section 3.2 introduces our machine learning methodology and presents the results. We compare approaches in Section 3.3. In Section 4, we explore heterogeneity in realizations rates and discuss the policy implications of our results. Section 5 concludes.

4. Cicala (2017) implements a variant on this methodology, using random forests rather than LASSO, in the context of electricity market integration. Varian (2016) provides an overview of causal inference targeted at scholars familiar with machine learning. He proposes using machine learning techniques to predict counterfactuals in a conceptually similar manner, although he does not implement his approach in an empirical setting.

2 Context and data

Existing engineering estimates suggest that commercial buildings, including schools, may present important opportunities to increase energy efficiency. For example, McKinsey & Company, who developed the iconic global abatement cost curve (see McKinsey & Company (2009)), note that buildings account for 18 percent of global emissions and as much as 30 percent in many developed countries. In turn, commercial buildings account for 32 percent of building emissions, with residential buildings making up the balance. Opportunities to improve commercial building efficiency primarily revolve around lighting, office equipment, and HVAC systems.

Commercial buildings such as schools, which are not operated by profit-maximizing agents, may be less likely to take advantage of cost-effective investments in energy efficiency, meaning that targeted programs to encourage investment in energy efficiency may yield particularly high returns among these establishments. On the other hand, schools are open fewer hours than many commercial buildings, so the returns may be lower.

We analyze schools that participated in Pacific Gas and Electric Company's (PG&E's) energy efficiency programs. School districts identified opportunities for improvements at their schools and then applied to PG&E for rebates to help cover the costs of qualifying investments. In California, utility energy efficiency programs are funded by a small adder on electricity and gas customer bills, which provides over $1 billion per year for programs across the residential, commercial and industrial sectors. Rates for California utilities have been "decoupled" for a number of years, meaning that investments in energy efficiency do not lower their revenue. The California Public Utility Commission oversees the utility energy efficiency programs to try to ensure that the utilities are providing incentives for savings that would not have been realized absent the utility program.

Energy efficiency retrofits for schools gained prominence in California with Proposition 39, which voters passed in November 2012. The proposition closed a corporate tax loophole and devoted half of the revenues to reducing the amount public schools spend on energy, largely through energy efficiency retrofits. Over the first three fiscal years of the program, the California legislature appropriated $1 billion to the program (California Energy Commission (2017)). This represents about one-third of what California spent on *all* utility-funded energy efficiency programs (ranging from low-interest financing to light bulb subsidies to complex industrial programs) and about 5 percent of what utilities nationwide spent on energy efficiency over the same time period (Barbose et al. (2013)). The upgrades we study in this paper largely predate the investments financed through Proposition 39, but are similar to the later projects, making our results relevant to expected energy savings from this large public program.

Methodologically, schools provide a convenient laboratory in which to isolate the impacts of energy efficiency. School buildings are all engaged in relatively similar activities, are subject to the same wide-ranging trends in education, and are clustered within distinct neighborhoods and towns.

Other commercial buildings, by contrast, can house anything from an energy intensive data center that operates around the clock to a church that operates very few hours per week. Finally, given the public nature of schools, we are able to assemble relatively detailed data on school characteristics and recent investments.

Most of the existing empirical work on energy efficiency focuses on the residential sector. There is little existing work on energy efficiency in commercial buildings. Kahn, Kok, and Quigley (2014) provide descriptive evidence on differences in energy consumption across one utility's commercial buildings as a function of various observables, including incentives embedded in the occupants' leases, age, and other physical attributes of the buildings. In other work, Kok and co-authors analyze the financial returns to energy efficiency attributes, though many of the attributes were part of the building's original construction and not part of deliberate retrofits, which are the focus of our work (Kok and Jennen (2012) and Eichholtz, Kok, and Quigley (2013)).

There is also a large grey literature evaluating energy efficiency programs, mostly through regulatory proceedings. Recent evaluations of energy efficiency programs for commercial customers, such as schools, in California find that actual savings are around 50 percent of projected savings for many efficiency investments (Itron (2017a)) and closer to 100 percent for lighting projects (Itron (2017b)). The methodologies in these studies combine process evaluation (e.g., verifying the number of light bulbs that were actually replaced) with impact evaluation, although the latter do not use meter-level data and instead rely on site visits by engineers to improve the inputs to engineering simulations. Recent studies explore the advantages of automating energy efficiency evaluations exploiting the richness of smart meter data and highlight the potential for the use of machine learning in this area (Granderson et al. (2017)). In this paper, we implement one of the first quasi-experimental evaluations of energy efficiency upgrades outside the residential sector.

2.1 Data sources

We use data from several sources. In particular, we combine high-frequency electricity consumption and account information with data on energy efficiency upgrades, school characteristics, community demographics, and weather. We obtain hourly interval electricity metering data for the universe of public K-12 schools in Northern California served by PG&E. The data begin in January 2008, or the first month after the school's smart meter was installed, whichever comes later.[5] 20 percent of the schools in the sample appear in 2008; the median year schools enter the sample is 2011. The data series runs through 2014.

In general, PG&E's databases link meters to customers for billing purposes. For schools, this

5. The raw PG&E interval data recorded consumption information every 15 minutes; we collapse these data to the hourly level because 15-minute level intervals are often missing. We take the average electricity consumption as representative, even if some of the 15-minute intervals are missing, to obtain a more balanced panel. Similarly, we interpolate consumption at a given hour if consumption at no more than two consecutive hours is missing.

creates a unique challenge: in general, school bills are paid by the district, rather than individual school. In order to estimate the effect of energy efficiency investments on electricity consumption, we required a concordance between meters and schools. We developed a meter matching process in parallel with PG&E. The final algorithm that was used to match meters to schools was implemented as follows: first, PG&E retrieved all meters associated with "education" customers by NAICS code.[6] Next, they used GPS coordinates attached to each meter to match meters from this universe to school sites, using school location data from the California Department of Education. This results in a good but imperfect match between meters and schools: in some cases, multiple school sites match to one or more meters. This can often be resolved by hand, and was wherever possible, but several "clusters" remain. We use only school-meter matches that did not need to be aggregated. Our final sample includes 1,870 schools.

The PG&E data also describe energy efficiency upgrades as long as the district applied for rebates from the utility.[7] 2,484 upgrades occurred at 911 schools between January 2008 and December 2014. For each energy efficiency measure installed, our data include the measure code, the measure description[8], a technology family (e.g., "HVAC", "Lighting", "Food service technology"), the number of units installed, the installation date, the expected lifetime of the project, the engineering-estimate of expected annual kWh savings, the incremental measure cost, and the PG&E upgrade incentive received by the school.[9] Many schools undertake multiple upgrades, either within or across categories. The engineering estimate of expected annual kWh savings and expected lifetime of the project are developed by the utility, which faces a strong incentive to increase estimated savings in order to demonstrate a successful program. In principle, regulatory oversight helps keeps the incentives to overstate savings in check, although in practice, the regulator has limited scope to penalize the utility for this.

We also obtain school and school-by-year information from the California Department of Education on academic performance, number of students, the demographic composition of each school's students, the type of school (i.e., elementary, middle school, high school or other) and location. We matched schools and school districts to Census blocks in order to incorporate additional neighborhood demographic information, such as racial composition and income. Finally, we obtain information on whether school district voters had approved facilities bonds in the two to five years

6. PG&E records a NAICS code for most customers in its system; this list of education customers was based on the customer NAICS code.

7. Anecdotally, the upgrades in our database are likely to make up a large share of energy efficiency upgrades undertaken by schools. PG&E reports making concerted marketing efforts to reach out to districts to induce them to make these investments; districts often lack funds to devote to energy efficiency upgrades in the absence of such rebates.

8. One example of a lighting measure description from our data: "PREMIUM T-8/T-5 28W ELEC BALLAST REPLACE T12 40W MAGN BALLAST-4 FT 2 LAMP"

9. We have opted not to use the cost data as we were unable to obtain a consistent definition of the variables related to costs.

before retrofits began at treated schools.[10]

We download hourly temperature data from 2008 to 2014 from over 4,500 weather stations across California from MesoWest, a weather data aggregation project hosted by the University of Utah.[11] We match school GPS coordinates provided by the Department of Education with weather station locations from MesoWest to pair each school with its closest weather station to create a school-specific hourly temperature record.

2.2 Summary statistics

Table 1 displays summary statistics for the data described above, across schools with and without energy efficiency projects. Of the 1,870 schools in the sample, 912 undertook at least one energy efficiency upgrade. There are 958 "untreated" schools that did not install any energy efficiency upgrades during our sample period. Our main variable of interest is hourly electricity consumption. We observe electricity consumption data for the average school for a three-year period. For schools that are treated, expected energy savings are almost 30,000 kWh, or approximately 5 percent of average annual electricity consumption.[12]

[Table 1 and Figure 1 about here]

Table 1 highlights measurable differences between treated and untreated schools. Treated schools consume substantially more electricity, appear in our sample earlier, are larger, and tend to be located to the southeast of untreated schools.

2.3 Trends in school characteristics

Because schools are different on a range of observable characteristics, and because these indicators may be correlated with electricity usage, it is important that we consider selection into treatment as a possible threat to econometric identification in this setting. One potential reassuring feature, highlighted by Figure 1, is that, in spite of the measurable differences across schools, there is substantial geographical overlap between them.

Because we have repeated observations for each school over time, we will employ a panel fixed effects approach, meaning that level differences alone do not constitute threats to identification. For our results to be biased, there must be *time-varying* differences between treated and untreated schools which correlate with the timing of energy efficiency upgrades. In order to examine the extent to which this is occurring, we examine patterns in five key school characteristics across treated and

10. Bond data are from EdSource (edsource.org).

11. We performed our own sample cleaning procedure on the data from these stations, dropping observations with unreasonably large fluctuations in temperature, and dropping stations with more than 10% missing or bad observations. The raw data are available with a free login from http://mesowest.utah.edu/.

12. We do not summarize expected savings in Table 1, as all untreated schools have expected savings of zero.

untreated schools over time using an event study specification. In particular, we examine the number of enrolled students, number of staff members, and the percentage of students performing "proficient" or better – the state standard – on California's Standardized Testing and Reporting (STAR) math and English/language arts exams, and energy consupmtion. Our estimating equation is:

$$Y_{it} = \sum_{y=-3}^{5} \beta^y \mathbf{1}[\text{Year to upgrade } = y]_{it} + \alpha_i + \gamma_t + \varepsilon_{it} \tag{2.1}$$

where Y_{it} is our outcome of interest for school i in year t, $\mathbf{1}[\text{Year to upgrade } = y]_{it}$ is an indicator defining "event time," such that $y = 0$ is the year of the energy efficiency upgrade, $y - 3$ is 3 years prior to the upgrade, and $y + 5$ is 5 years after the upgrade, etc. α_i is a school fixed effect, γ_t is a year fixed effect, and ε_{it} is an error term, which we cluster at the school level.[13] Figure 2 displays the results of this exercise.

[Figure 2 about here]

Across the four demographic variables, we see that treated and untreated schools are behaving similarly before and after energy efficiency upgrades. The relatively flat pre- and post-treatment trends is evidence in favor of our identifying assumption that treated and untreated schools were and would have remained on parallel trends in the absence of energy efficiency upgrades. In particular, the results on the number of students and number of staff suggest that treated schools did not grow or shrink substantially at the same time as they installed energy efficiency upgrades, and the test score results provide evidence that schools' instructional quality did not change dramatically around energy efficiency upgrades. We can rule out even small changes in all four variables; we find precisely-estimated null results.

The final panel of Figure 2 provides suggestive evidence that treated and untreated schools had similar trends in energy consumption prior to energy efficiency upgrades. Furthermore, we find that these upgrades are associated with a marked decline in energy consumption at treated schools. This panel lends further support to our assumption that treated and untreated schools would have remained on similar trajectories in the absence of energy efficiency upgrades, and suggests that energy efficiency upgrades caused a substantial reduction in energy use.

3 Empirical strategy and results

In this section, we describe our empirical approach and present results. We begin with a standard panel fixed effects strategy. Despite including a rich set of fixed effects in all specifications, we demonstrate that this approach is highly sensitive to both specification and outliers. We proceed

13. Because we have richer data on electricity consumption, we include a school-by-hour-of-day fixed effect rather than a school fixed effect in this final regression.

by implementing a machine learning methodology, wherein we generate school-specific models of electricity consumption to construct counterfactual electricity use in the absence of energy efficiency upgrades. We demonstrate that this method is substantially less sensitive to specification and sample restrictions than our regression analysis, and enables us to select among the millions of possible covariates in a systematic way.

3.1 Panel fixed effects approach

3.1.1 Methodology

The first step of our empirical analysis is to estimate the causal impact of energy efficiency upgrades on electricity consumption. In an ideal experiment, we would randomly assign upgrades to some schools and not to others. In the absence of such an experiment, we begin by turning to standard quasi-experimental methods. We are interested in estimating the following equation:

$$Y_{ith} = \beta D_{it} + \alpha_{ith} + \varepsilon_{ith} \tag{3.1}$$

where Y_{ith} is energy consumption in kWh at school i on date t during hour-of-day h. Our treatment indicator, D_{it}, is a dummy indicating that school i has undertaken at least one energy efficiency upgrade by date t.[14] The coefficient of interest, β, can be interpreted as the average savings in kWh/hour at a treated school. α_{ith} represents a variety of possible fixed effects approaches. Because of the richness of our data, we are able to include many multi-dimensional fixed effects, which non-parametrically control for observable and unobservable characteristics that vary across schools and time periods. Finally, ε_{ith} is an error term, which we cluster at the school level to account for arbitrary within-school correlations.[15]

We present results from several specifications with increasingly stringent controls. In our most parsimonious specification, we control for school-by-hour-of-day fixed effects, accounting for hour-specific time-invariant characteristics at each school. Our preferred specification includes school-by-hour-by-month-of-year fixed effects, to control for differential patterns of electricity consumption across schools, and month-of-sample fixed effects, to control for common shocks or time trends in energy consumption. As a result, our econometric identification comes from within-school-by-hour-

14. Though schools can and do undertake multiple upgrades, we use a binary treatment indicator here due to concern about mismeasurement of treatment dates. When we instead define D_{it} as the cumulative number of upgrades undertaken by school i by time t, we find smaller realization rates, further supporting our conclusion that energy efficiency upgrades deliver less than the expected savings. Discussions with the utility confirmed that there is substantial heterogeneity on how accurately the dates are recorded.

15. To speed computation time, the regressions presented in the paper were estimated by first collapsing the data to the school-by-month-of-sample-by-hour-of-day level. This collapse averages over identifying variation driven by different patterns across days of the week, but enables us to more easily include month-of-sample and school-hour-specific fixed effects. After collapsing the data, we re-weight our regressions such that we recover results that are equivalent to first order to our estimates on the disaggregated data. Results from the uncollapsed data are virtually the same and available upon request.

month-of-year and within-month-of-sample differences between treated and untreated schools.

Realization rates In addition to estimating impacts of energy efficiency upgrades on energy consumption, we compare these estimates to average *ex ante* estimates of expected savings. We follow the existing energy efficiency literature in calculating realization rates.[16] Specifically, we calculate the realization rate as $\hat{\beta}$ divided by the average expected savings for upgrades in our sample. To ensure that the average savings are properly weighted to match the relevant regression sample, we compute these average savings by regressing expected savings for each school at a given time t (equal to savings by time t for treated schools in the post-treatment period, and zero otherwise) on the treatment time variable and the same set of controls and fixed effects as its corresponding regression specification. If our *ex post* estimate of average realized savings matches the *ex ante* engineering estimate, we will estimate a realization rate of one. Realization rates below (above) one imply that realized savings are lower (higher) than expected savings.

3.1.2 Results

Table 2 reports results from estimating Equation (3.1) using five different sets of fixed effects. We find that energy efficiency upgrades resulted in energy consumption reductions of between 1.3 and 3.5 kWh/hour. These results are highly sensitive to the set of fixed effects included in the regression. Using our preferred specification, Column (5) in Table 2, which includes school-by-hour-by-month-of-year and month-of-sample fixed effects, we find that energy efficiency upgrades caused a 1.81 kWh/hour reduction in energy consumption at treated schools. In column (6), we also control for temperature, and find a 1.75 kWh/hour reduction in energy consumption, and a realization rate of 0.74. These results are all precisely estimated; all energy savings estimates are statistically significant at the 1 percent level.[17]

[Table 2 about here]

Using this panel fixed effects approach, we find evidence that energy efficiency upgrades reduced school electricity consumption. However, these upgrades appear to under-deliver relative to *ex ante* expectations. In all specifications, we find realization rates below one: our estimated realization rates range from 0.54 to 0.90. This suggests that energy savings in schools are not as large as expected. In our most comprehensive specification, which includes a temperature control, the realization rate is 0.74. In this case, we cannot reject a rate of one.

16. Davis, Fuchs, and Gertler (2014), Fowlie, Greenstone, and Wolfram (2018), Levinson (2016b), Kotchen (2017), Novan and Smith (2018), and Allcott and Greenstone (2017) all use this method.

17. In Appendix Table A.1, we present standard errors using two-way clustering on school and month of sample, allowing for arbitrary dependence within schools and across schools within a time period. The results remain highly statistically significant using these alternative approaches.

3.1.3 Panel fixed effects robustness

Trimming We subject our panel fixed effects approach to a number of standard robustness checks. We begin by examining the sensitivity of our estimates to outliers. This is particularly important in our context, because we run our main specifications in levels to facilitate the computation of realization rates. Table 3 repeats the estimates from Table 2 with three different approaches to removing outliers. In Panel A, we trim observations below the 1st or above the 99th percentile of energy consumption. Doing so reduces the point estimates dramatically. We now estimate savings between 0.28 kWh/hour and 2.49 kWh/hour. This trimming also has substantial impacts on our realization rate estimates, which now range from 0.11 to 0.59.

In Panel B, we instead trim schools below the 1st and above the 99th percentile in terms of expected savings. We implement this trim because expected savings has an extremely skewed distribution in our sample.[18] We find that the results are less sensitive to this trim than the trim in Panel A; we now estimate point estimates between 1.02 kWh/hour and 3.27 kWh/hour, and realization rates between 0.44 and 0.85.

In Panel C, we implement both trims together, and the results are similar to those in Panel A. We again find much lower point estimates (ranging from 0.26 kWh/hour to 2.43 kWh/hour) and realization rates (ranging from 0.11 to 0.63) than in the full sample.

Overall, the panel fixed effects estimates are extremely sensitive to both specification and to outliers in the sample. This is concerning from a policy perspective; realization rates between 0.54 and 0.90 have substantially different implications than rates between 0.11 and 0.63, and is also cause for concern about the performance of the panel fixed effects estimator in this context. Controlling for temperature in specification (6) helps mitigate the effects of trimming, but the results remain sensitive to outliers, with estimated realization rates moving between 0.74 with no trimming to 0.53 when trimming outlier observations.

It bears pointing out that there are many possible variants on the panel fixed effects design (see for example the matching approach from Ferraro and Miranda (2017) and Cicala (2015) or the Abadie, Diamond, and Hainmueller (2010) synthetic control method). Given the richness of our data, we also have a great deal of flexibility in our choice of control variables, fixed effects, and functional form.[19] The results presented above come from a fairly standard parsimonious specification. In order to add additional controls in an algorithmic fashion, we now turn to a

18. The median project was expected to save 16,663 kWh, while the average project was expected to save 46,050 kWh. We believe some of this to be measurement error; five percent of schools in the sample which are expected to reduce their energy consumption by 50 percent through energy efficiency upgrades, which seems unrealistic.

19. As one variant on our main approach, we conduct a limited nearest neighbor matching exercise, in which we use observable characteristics of treated schools to find similar untreated schools. Appendix Table A.2 displays the results, using three different candidate control groups: all untreated schools; schools in the same district as the treated school only; and schools in other districts only. These results are highly sensitive to specification and the selected control group. As a result, we turn to the machine learning approach, which is in the spirit of the synthetic control method, to select covariates.

machine learning approach.

3.2 Machine learning approach

Even with a large set of high-dimensional fixed effects, the standard panel approach performs poorly on basic robustness tests, and is extremely sensitive to specification. A natural next step would be to add additional controls. However, given the size of the dataset, a researcher interested in capturing heterogeneity could interact several variables with school and hour-of-day, generating millions of candidate covariates. This makes the process of model selection computationally expensive and ad hoc. In order to address some of these issues more systematically, we use a machine learning approach that leverages the richness of the data.

3.2.1 Methodology overview

We use machine learning methods to generate counterfactual models of energy consumption in the absence of energy efficiency upgrades. Machine learning is particularly well-suited to constructing counterfactuals, since the goal of building the counterfactual is not to isolate the effect of any particular variable, but rather to generate a good overall prediction. Because machine learning methods do model selection via algorithm, including cross-validation, these models tend to generate better out-of-sample predictions than models chosen by researchers (Abadie and Kasy (2017)). These methods also enable researchers to allow for a substantially wider covariate space than would be feasible with trial-and-error. These features make machine learning methods particularly attractive for applied microeconomists. Our methodology, which embeds machine learning methods in a traditional panel fixed effects approach, proceeds in two steps. Figure 3 provides an overview of these steps.

[Figure 3 about here]

In a **first step**, we use machine learning tools to create unit-specific models of an outcome of interest. We train these models using pre-treatment data only, which ensures that variable selection is not confounded by structural changes that occur in the post-treatment period. We then use these models to create (fully out-of-sample) predictions of our outcome of interest in the post-treatment period. We compare the machine learning predictions to real data to compute prediction errors for each unit.

In a **second step**, we leverage the fact that some schools are treated and some are not, to estimate pooled panel fixed effects regressions with these prediction errors as the dependent variable. This combination of machine learning methods with panel fixed effects approaches enables us to control for confounding trends and address other possible threats to identification. We leverage

within-unit within-time-period variation for identification while parsimoniously choosing among millions of potential control variables in a highly flexible and computationally feasible way.[20]

Our regression specification is analogous to our panel fixed effects model, described in Equation (3.1), but we now use the prediction error as the dependent variable:

$$Y_{ith} - \hat{Y}_{ith} = \beta D_{it} + \alpha_{ith} + \gamma \text{posttrain}_{ith} + \varepsilon_{ith}, \tag{3.2}$$

where α_{ith} and ε_{ith} are defined as in Equation (3.1), $\hat{Y}_{ith}$ is the prediction in kWh from step one and posttrain_{ith} is a dummy, equal to one during the out-of-sample prediction period. We include this dummy to account for possible bias in the out-of-sample predictions, by re-centering prediction errors in the untreated schools around zero.[21] We cluster our standard errors at the school level. Because we care about the expectation of the prediction, rather than the prediction itself, our standard errors are unlikely to be substantially underestimated by failing to explicitly account for our forecasted dependent variable.[22]

Identification As with the standard panel fixed effects approach, the identifying assumption is that, conditional on control variables, energy consumption at treated and untreated schools would have been trending similar;y in the absence of treatment. In this specification, we require treated and untreated schools to be trending similarly in *prediction errors*, rather than in energy consumption. This is analogous to having included a much richer set of control variables on the right-hand side of our regression. In a sense, the machine learning methodology enables us to run a much more flexible model in a parsimonious, computationally tractable, and systematic way.

It is important to note, however, that our machine learning approach —just like the panel fixed effects approach— is not immune from bias stemming from energy consumption changes that coincide directly with the subsidized energy efficiency upgrades. If a school undertakes additional energy-saving behaviors or unsubsidized upgrades at the same time as an energy efficiency upgrade in our sample, we will overestimate energy savings and the resulting realization rates will be over-

20. Machine learning methods have become increasingly popular in economics. Athey (2017) and Mullainathan and Spiess (2017) provide useful overviews. Our paper extends a strand of this literature which combines machine learning techniques with quasi-experimental econometric methods. This includes McCaffrey, Ridgeway, and Morral (2004), who propose a machine learning based propensity score matching method; Wyss et al. (2014), who force covariate "balance" by directly including balancing constraints in the machine learning algorithm used to predict selection into treatment; and Belloni, Chernozhukov, and Hansen (2014) propose a "double selection" approach, using machine learning to both predict selection into treatment as well as to predict an outcome, using both the covariates that predict treatment assignment and the outcome in the final step. In our panel data context, predicting selection into treatment is unnecessary, as this is absorbed by unit fixed effects. Our paper is most similar in spirit to Athey et al. (2017), in which the authors propose a matrix completion method for estimating counterfactuals in panel data, but differs in that we can account for a large amount of heterogeneity in a computationally tractable way.

21. As shown in Panel D of Figure 4 below, these prediction errors are centered around zero in our application, so in practice this has a minimal impact on the results. However, this correction could be important in other settings.

22. Other papers which employ machine learning methods treat standard errors similarly, e.g. Cicala (2017) and Deryugina et al. (2019).

estimates. Any remaining positive selection into treatment, for instance, based on the expected size of the treatment effect, will bias our estimates away from zero, leading us to estimate energy efficiency savings and realization rates that are more favorable. For a confounder to bias our results towards zero, a school would have to increase energy use at the same time as our upgrades. We provide suggestive evidence against this in Figure 2, where we show that school size, number of staff, and test scores do not change dramatically around the time of upgrade. This does not rule out the possibility of dramatic changes in energy usage that were coincident with energy efficiency upgrades, but it does appear unlikely that major schooling changes are driving our results. As a result, we believe that our approach is likely to deliver, if anything, estimates that are biased away from zero, providing an upper bound on the effectiveness of energy efficiency upgrades.

We continue by providing a more thorough discussion of our machine learning methodology and describing the results.

3.2.2 Step 1: Predicting counterfactuals

In the first step, we use machine learning to construct school-by-hour-of-day specific prediction models. For treated schools, we define the pre-treatment period as the period before any intervention occurs. For untreated schools, we randomly assign a "treatment date," which we use to define the "pre-treatment" period.[23] We train these models using pre-treatment data only, as described above.[24]

There are many possible supervised machine learning methods that researchers could use in this step. In our baseline approach, we use the Least Absolute Shrinkage and Selection Operator (LASSO), a form of regularized regression, to generate a model of energy consumption at each school.[25] We allow the LASSO to search over a large set of potential covariates, including the day of the week, a holiday dummy, a month dummy, a temperature spline, the maximum and minimum temperature for the day, and interactions between these variables. Because we are estimating school-hour-specific models, each covariate is also essentially interacted with a school fixed effect and an hour fixed effect—meaning that the full covariate space includes over 12,000,000 candidate

23. We randomly assign this date between the 20th and 80th percentile of in-sample calendar dates in order to have a more balanced number of observations in the pre- and post-sample, similar to that in the treated schools.

24. As an example, suppose that we observe an untreated school that we observe between 2009 and 2013. We randomly select a cutoff date for this particular school, e.g., March 3, 2011, and only use data prior to this cutoff date when generating our prediction model. For a treated school with a treatment date of July 16, 2012, we use only data prior to this date while to generate the prediction models.

25. We also consider variants on the LASSO and two random forest approaches, as well as alternative tuning parameters. We use the correlation between the predicted and actual energy consumption for untreated schools in the post-training period as an out-of-sample check on the performance of these different models. Table A.3 displays the results of this exercise, showing the distribution of correlations between data and predictions across these six methods. Our chosen method, including basic variables and untreated schools, and using `glmnet's` default tuning parameter, performs slightly better than the other options. We also explore results using these different models in Appendix Figure A.1, which shows that hour-specific treatment effects are robust to the choice of method.

variables.[26,27] In addition to these unit-specific variables, we also include consumption at untreated schools as a potential predictor, in the spirit of the synthetic control literature (Abadie, Diamond, and Hainmueller (2010)). The LASSO algorithm uses then cross-validation to parameterize the degree of saturation of the model and pick the variables that are included.[28]

Validity checks We perform several diagnostic tests to assess the performance of our predictions. Figure 4 presents four such checks. First, Panel A plots the number of selected covariates for each model against the size of the pre-treatment sample. LASSO penalizes extraneous variables, meaning that the optimal model for any given school will not include all of the candidate regressors.[29] Though the LASSO typically selects fewer than 100 variables, the joint set of variables selected across all schools and hours covers the majority of the candidate space (a total of 1,149 variables are selected), highlighting the importance of between-school heterogeneity.

[Figure 4 about here]

We can also inspect the selected covariates individually. As an illustration, Panel B of Figure 4 shows the coefficient on the holiday dummy (and its interactions) in each school-hour-specific prediction model.[30] We find that, across models, holidays are negatively associated with energy consumption. This suggests that the LASSO-selected models reflect real-world electricity use. We also find substantial heterogeneity across schools: each of the candidate holiday variables is selected at least once, but the median school has no holiday variable, highlighting the importance of data-driven model selection.

Panel C of Figure 4 shows the variables selected by each of the school-hour models for treated and untreated schools separately. Nearly all of the models include an intercept, and around 70 percent of the models include consumption from at least one untreated school; the median school-hour model includes ten such covariates. Month and temperature variables are each included in nearly half of the models. Several models also include interactions between temperature and

26. To make the approach computationally tractable, we estimate a LASSO model one school-hour at a time.

27. Note that we do not include time trends in the prediction model, because we are generating predictions substantially out of sample and these trends could dramatically drive predictions. The underlying assumption necessary for the predictions to be accurate is that units are in a relatively static environment, at least on average, which seems reasonable in this particular application.

28. We use the package `glmnet` in R to implement the estimation of each model. To cross-validate the model, the algorithm separates the pre-treatment data (from one school at a time) into "training" and "testing" sets. The algorithm finds the model with the best fit in the training data, and then tests the out-of-sample fit of this model in the testing set. We tune the `glmnet` method to perform cross-validation using a block-bootstrap approach, in which each week is considered to be a potential draw. This allows us to take into account potential autocorrelation in the data.

29. The LASSO performs best when the underlying DGP is sparse (Abadie and Kasy (2017)). We find evidence in favor of this in our empirical context, as the number of chosen regressors does not scale linearly with the size of the training set.

30. We define "holidays" to include major national holidays, as well as the Thanksgiving and winter break common to most schools. Unfortunately, we do not have school-level data for the exact dates of summer vacations, although the seasonal splines should help account for any long spells of inactivity at the schools.

weekday dummies. This again demonstrates the substantial heterogeneity in prediction models across schools, and suggests that our machine learning method yields counterfactual predictions that are substantially more flexible than their traditional panel fixed effects analogue, wherein we would estimate the same covariates for each unit.

Finally, we can perform a fully out-of-sample test of our approach by inspecting prediction errors at untreated schools in the post-treatment period. Because these schools do not experience energy efficiency upgrades, these prediction errors should be close to zero. Panel D of Figure 4 plots the distribution of average out-of-sample prediction error for each school-hour, trimming the top and bottom 1 percent. As expected, this distribution is centered around zero. Taken together, these four checks provide evidence that the machine learning approach is performing well in predicting schools' electricity consumption, even out-of-sample.

3.2.3 Step 2: Panel regressions with prediction errors

We now regress the prediction errors from the machine learning model on a treatment indicator and the rich set of fixed effects we use in the earlier panel fixed effects approach. Table 4 reports results from estimating Equation (3.2) for five different fixed effects specifications. We find that energy efficiency upgrades resulted in energy consumption reductions of between 2.2 and 4.2 kWh/hour. In our preferred specification (Column (5)), which includes school-by-hour-by-month and month-of-sample fixed effects, we find that energy efficiency upgrades reduced electricity use by 2.61 kWh/hour in treated schools relative to untreated schools. These results are both larger and more stable across specifications than the panel fixed effects results above, and are highly statistically significant.[31]

[Table 4 about here]

We again compare these results to the *ex ante* engineering estimates to form realization rates. Our estimated realization rates range from 0.70 to 1.01. These realization rates are statistically different than zero and larger than the estimates from our panel fixed effects approach. Some of the specifications imply that realized savings were in line with expected savings, with our preferred specification implying a realization rate of 78 percent.

3.2.4 Machine learning robustness

Trimming As with the panel fixed effects approach, we test the extent to which our machine learning results vary as we exclude outlying observations. Table 4 presents the results of this exercise. In Panel A, we drop observations that are below the 1st or above the 99th percentile of

31. In Appendix Table A.4, we present results two-way clustering on school and month of sample. The results remain highly statistically significant using these alternative approaches.

the dependent variable – now defined as prediction errors in energy consumption. Unlike in the panel fixed effects approach, we find that this trimming has very limited impacts on the results. We now find point estimates ranging from -3.68 kWh/hour to -2.20 kWh/hour, and accompanying realization rates ranging from 0.89 to 0.65. These are similar to our estimates in Table 4. In Panel B, we again trim schools with expected savings below the 1st or above the 99th percentile. We find that this, too, neither meaningfully alters our point estimates nor our realization rates, which now range from -3.93 kWh/hour to -1.98 kWh/hour and 1.02 to 0.66, respectively. Finally, in Panel C, we trim on both dimensions, and again find remarkably stable point estimates and realization rates, ranging from -3.55 to -2.10 kWh/hour and 0.94 to 0.64. While the panel fixed effects results displayed in Table 3 were highly sensitive to these trimming approaches, the machine learning results are quite stable.

Alternative prediction approaches How sensitive are our results to our use and implementation of the LASSO algorithm? Depending on the underlying data, different algorithms may be more effective than others (Mullainathan and Spiess (2017)). As described in Section 3.2.1, the LASSO appears to generate well-behaved models. Furthermore, we find similar out-of-sample prediction effectiveness in untreated schools across our choice of tuning parameters and potential covariates, as well as when we train our models using a random forest algorithm rather than a LASSO algorithm. Similarly, we find that our main results look remarkably similar across these techniques. Appendix Table A.5 shows the results where we estimate (3.1) with different prediction algorithm approaches. We find energy savings between 2.44 kWh per hour and 2.64 kWh per hour. Using our preferred LASSO approach (Column (4)), we estimate savings of 2.61 kWh per hour. These estimates translate into realization rates of 0.73, 0.79, and 0.78, respectively.[32] These estimates are generally not statistically distinguishable, and our preferred approach lies in the middle of the range of estimated realization rates, suggesting that the machine learning approach is not highly sensitive to our chosen prediction method.

3.3 Comparing approaches

In contrast with the standard panel fixed effects approach, our machine learning method delivers results that are larger and substantially less sensitive to both specification and sample selection. This highlights one advantage of using machine learning approaches in panel settings: by controlling for confounding factors using a flexible data-driven approach, this method can produce results that are more robust to remaining researcher choices.

We explore this result further in Figure 5, which shows the distribution of estimated realization

32. Appendix Figure A.1 shows hour-specific treatment effects for all of the machine learning methods shown here. Both the hourly patterns and the levels are very similar across methods.

rates across several specifications and samples.[33] Notably, the policy implications from the different panel fixed effects estimates vary widely, and are centered around a 50 percent realization rate, whereas the estimates using the machine learning approach are more stable around realization rates closer to 100 percent.

[Figure 5 about here]

One potential criticism of our panel approach is that it is not sufficiently rich. For the purposes of comparison, we estimate additional specifications in which, in addition to the fixed effects we include above, we add temperature interacted with a school fixed effect: an extremely flexible temperature control. We estimate these regressions on the samples described above, and add these additional results to Figure 5. Controlling for temperature does reduce the sensitivity of the panel fixed effects regressions, but the resulting estimates remain more variable than those estimated using the machine learning approach.

While researchers could attempt a variety of alternative specifications in an ad-hoc way in order to reduce sensitivity to specification and sample, this approach is impractical with high-frequency datasets. With over 12,000,000 possible covariates to choose from, doing model selection by hand is computationally expensive and arbitrary.[34] In contrast, our machine learning approach enables researchers to perform model selection in a flexible yet systematic way, while maintaining the identifying assumptions needed for causal inference in a standard panel fixed effects approach.

4 Policy implications

Our preferred estimates imply that energy efficiency upgrades in public schools only delivered 78 percent of expected savings. What other lessons can we learn from the data? What are the cost-benefit implications of this finding?

4.1 Heterogeneity and targeting

Beyond estimating average realization rates, understanding whether these rates vary based on observable characteristics of upgrades or treated schools may be informative for policymakers deciding which upgrades to subsidize and which schools to target.[35]

33. The results include five specifications per method (the ones in the main Tables (3.1) and (3.2)). We estimate each of the five specifications on five different samples: no trimming, trimming the top and bottom 1 and 2 percent of observations within each school, trimming the schools with the smallest and largest 1 percent of interventions, and a combination of 1 percent trimming for each school combined with removing schools with small and large interventions. Each resulting kernel density is composed of total of 25 estimates.

34. Given that we have an unbalanced panel, in which some schools are observed for longer periods than others, it is also unclear that saturating the model equally across schools is necessarily the best strategy.

35. There can also be heterogeneity in the timing of savings. Because our focus in this paper is on realization rates, which are determined by overall savings, we do not focus here on heterogeneity of treatment effects by time. As

Given the richness of our electricity consumption data, we start by estimating school-specific treatment effects, as a precursor to determining what drives heterogeneity in realization rates.[36] These estimates should not be taken as precise causal estimates of savings at any given school, but rather as an input to projecting heterogeneous estimates onto school-specific and intervention-specific covariates for descriptive purposes.

To compute these school-specific estimates, we regress prediction errors in kWh on a school-specific dummy variable, equal to one during the post-treatment period (or, for untreated schools, the post-training period from the machine learning model), as well as school-by-hour-by-month fixed effects to control for seasonality. The resulting estimates represent the difference between pre- and post-treatment energy consumption at each individual school. We can then use these school-specific estimates to understand the distribution of treatment effects, and try to recover potential systematic patterns across schools.

Panel A of Figure 6 displays the relationship between these school-specific savings estimates and expected savings for treated schools. We find a positive correlation between estimated savings and expected savings, although there is substantial noise in the school-specific estimates. Once we trim outliers in expected savings, we recover a slope of 0.54. Panel B presents a comparison of the school-specific effects between treated and untreated schools. The estimates at untreated schools are much more tightly centered around zero, in line with Panel D of Figure 4. In contrast, the distribution of treated school estimates is shifted towards additional savings, consistent with schools having saved energy as a result of their energy efficiency upgrades. These results suggest that — in keeping with our main results — energy efficiency projects successfully deliver savings, although the relationship between the savings that we can measure and the *ex ante* predicted savings is noisy.

[Figure 6 about here]

We next try to project these school-specific estimates onto information that is readily available to policymakers, in an attempt to find predictors of higher realization rates. We do this by regressing our school-specific treatment effects onto a variety of covariates via quantile regression, in order to remove the undue influence of outliers in these noisy estimates.[37] We include one observation per

Borenstein (2002) and Boomhower and Davis (2017) point out, however, the value of energy savings varies over time. We also estimate hour-specific treatment effects, presented in Appendix Figure A.1, across several machine learning methods. We find evidence that the largest reductions occur during the school day, consistent with our results picking up real, rather than spurious, energy savings. This is suggestive that the reductions in our sample are happening at relatively high-value times, though peak power consumption hours in California occur between 4 and 8 PM, after the largest estimated reductions from the energy efficiency upgrades in our sample.

36. Naturally, the identifying assumptions required to obtain school-specific treatment effects are much stronger than when obtaining average treatment effects, as concurrent changes in consumption at each specific school will be confounded with its own estimated treatment effect (i.e., random coincidental shocks to a given school that might not confound an average treatment effect will certainly confound the school-specific estimate of that given school).

37. Note that we could also have used a quantile regression approach in our high-frequency data, which would assuage

treated school in our sample, and weight the observations by the length of the time series of energy data for each school.[38] We center all variables (except for dummy variables) around their mean and normalize by their standard deviation.

[Table 6 about here]

Table 6 presents the results of this exercise. Column (1) shows that the median realization rate for treated schools using this approach is close to 80 percent. Column (2) shows that median realization rates are larger for HVAC and lighting interventions (the most prevalent types of upgrades in our sample), although the estimates are very noisy. We add latitude, longitude and temperature in Column (3), but these are not significantly correlated with realization rates after controlling for the types of interventions. Columns (4)-(5) control for standardized values of yet more covariates, including the Academic Performance Index and the poverty rate. We find suggestive evidence that larger schools have higher realization rates, though we find no other statistically significant correlations between observable characteristics and realization rates.[39] These descriptive regressions should be interpreted with caution. These are cross-sectional estimates, and school size is likely correlated with a variety of other important factors, including intervention size. In Column (6), we look at the relationship between expected savings and realization rates directly. We find that after controlling for school size, larger interventions are assocaited with lower realization rates.

Ultimately, we uncover mostly noisy correlations between school characteristics and realization rates. This suggests that uncovering "low-hanging fruit" to improve the success of energy efficiency upgrades in this setting may be difficult. That said, several features of our setting make recovering this type of patterns challenging. Our sample of treated schools is relatively small—there are fewer than 1,000 observations in these quantile regressions, and each of the schools is subject to its own idiosyncrasies, leading to concerns about collinearity and omitted variables bias. It is possible that in samples with more homogeneous energy efficiency projects, and with a larger pool of treated units, it could be feasible to identify covariates that predict higher realization rates. This in turn could be used to inform targeting procedures to improve average performance.

4.2 Cost-benefit analysis

Our focus on this paper is on realization rates: we use schools as a useful empirical setting to estimate the effectiveness of energy efficiency upgrades in delivering predicted electricity savings. In particular, our interest lies in comparing *ex ante* engineering estimates of energy savings to *ex*

potential concerns about outliers. Because we rely on a large set of high-dimensional fixed effects for identification, however, this is computationally intractable.

38. Note that untreated schools are not included in these regressions, since they have no treatment effects by definition.

39. We explored a variety of other potential demographic variables, but we did not find any clear correlation with realization rates.

post realizations. We do not perform a cost-benefit analysis in this paper, which would require accounting for the full benefits of the energy efficiency upgrades as well as reliable cost data. Finally, if anything, the schools in our sample are already privately over-incentivized to invest in energy efficiency measures, because electricity prices in California are substantially higher than social marginal cost (Borenstein and Bushnell (2018)).[40]

First, energy efficiency upgrades may be associated with welfare benefits beyond reductions in electricity consumption. For example, consider an inefficient air conditioning unit that gets replaced with a more silent and efficient version that gets turned on more often, mitigating the negative impacts of high temperatures on human capital accumulation (e.g., Graff Zivin, Hsiang, and Neidell (2017)).[41] We provide suggestive evidence that energy efficiency upgrades do not improve standardized test scores in Figure 2, though test scores remain an imperfect proxy for human capital accumulation, and do not capture all possible non-energy benefits of energy efficiency improvements. Second, the data we obtained from PG&E do not contain comprehensive information on costs. In particular, the only cost information in our dataset is the "incremental measure cost," a measure of the difference in the cost of a "base case" appliance replacement versus an energy efficient version. We do not, however, have data on the total cost of the appliance replacement, nor on projected energy savings from the base case counterfactual, precluding us from a standard cost-benefit or return-on-investment analysis.

One potential way to assess the relevance of costs and benefits with our limited data is to use the CPUC's own cost-benefit analysis before approving an energy efficiency upgrade. In order for the CPUC to allow utilities to install subsidized energy efficiency upgrades, these upgrades must be determined to have a savings-to-investment ratio (SIR) of 1.05. That is, each upgrade must have *expected* savings of 1.05 times its investment cost – where expected savings are based on the same *ex ante* engineering estimates we exploit in this paper. We do not have (aggregated or micro) data on the SIRs in our sample, but in light of our central realization rate estimate of 78 percent, upgrades where the SIR was below 1.5 would not pass this CPUC test if the SIR were instead based on realized savings. However, without data on the average SIRs in this sample, we cannot directly comment on the proportion of measures that would no longer pass the cost-benefit test.

5 Conclusion

We leverage high-frequency data on electricity consumption and develop a machine learning method to estimate the causal effect of energy efficiency upgrades at K-12 schools in California. We use

40. Borenstein and Bushnell (2018) show that the social marginal costs of electricity generation in California are approximately 6 cents per kWh. Schools are typically on tariffs with rates between 8 and 12 cents per kWh.

41. Much of the existing literature which estimates the impacts of energy use on student achievement uses student-specific data (e.g., Park (2017) and Garg, Jagnani, and Taraz (2018)), to which we do not have access to. We leave these additional avenues for future work.

two main approaches to do this, both of which leverage cross-sectional and temporal variation to separate the causal effect of energy efficiency upgrades from other confounding factors. We begin with a panel fixed effects approach. Using this method, we estimate that energy efficiency upgrades saved 70 percent of *ex ante* estimated savings. However, these estimates are sensitive to specification and outliers, and range from 11 to 90 percent. Given the richness of our setting, there are millions of possible covariates we could include as controls. In order to parsimoniously select among these control variables, we implement a second approach, using tools from machine learning.

In our machine learning approach, we use untreated time periods in high-frequency panel data to generate school-specific predictions of energy consumption that would have occurred in the absence of treatment for both treated and untreated schools. We generate prediction errors by comparing these predictions to realized energy consumption, and estimate the causal effect of energy efficiency upgrades by estimating a panel fixed effects model using prediction errors as the outcome variable. This approach allows us to select among covariates in a parsimonious way, while still accounting for common shocks. Our approach is computationally tractable, and can be applied to a broad class of applied settings where researchers have access to relatively high-frequency panel data.

Using this method, we find that energy efficiency upgrades deliver 78 percent of *ex ante* expected savings on average. As compared to our panel fixed effects approach, we see that the machine learning approach delivers a much narrower range of estimates: energy efficiency upgrades deliver between 65 and 102 percent of expected savings, depending on outliers and specification. This highlights the potential benefits of using machine learning to select among a large set of exogenous control variables.

To draw policy implications, we explore heterogeneity in realization rates and discuss the cost-benefit of these upgrades. We find some evidence that HVAC and lighting upgrades outperform other upgrades. We attempt to use other information that is readily available to policymakers to predict which schools will have higher realization rates, but the results are noisy, and we ultimately find it difficult to identify school characteristics that systematically predict higher realization rates. This suggests that without collecting additional data, improving realization rates via targeting may prove challenging.

This paper extends the energy efficiency literature to a non-residential sector. We demonstrate that energy efficiency upgrades deliver lower savings than expected *ex ante*, although in several specifications we cannot reject full realization rates, and are able to reject some of the extremely low realization rates of the prior literature. These results have implications for policymakers and building managers deciding over a range of capital investments, and demonstrates the importance of real-world, *ex post* program evaluation in determining the effectiveness of energy efficiency. Beyond energy efficiency applications, we show how machine learning tools can help with specification choice, leading to results that are robust to the machine learning algorithm of choice, varying sets

of fixed effects, and the treatment of outliers.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program."

Abadie, Alberto, and Maximilian Kasy. 2017. "The Risk of Machine Learning." *Working paper.*

Allcott, Hunt, and Michael Greenstone. 2012. "Is there an energy efficiency gap?" *The Journal of Economic Perspectives* 6 (1): 3–28.

———. 2017. *Measuring the Welfare Effects of Residential Energy Efficiency Programs.* Technical report. National Bureau of Economic Research Working Paper No. 23386.

Athey, Susan. 2017. "Beyond prediction: Using big data for policy problems." *Science* 355 (6324): 483–485.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2017. *Matrix Completion Methods for Causal Panel Data Models.* Working Paper 1710.10251. arXiv.

Barbose, Galen L., Charles A. Goldman, Ian M. Hoffman, and Megan A. Billingsley. 2013. "The future of utility customer-funded energy efficiency programs in the United States: projected spending and savings to 2025." *Energy Efficiency Journal* 6 (3): 475–493.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls." *The Review of Economic Studies* 81 (2): 608–650.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350:1073–1076.

Boomhower, Judson, and Lucas Davis. 2017. "Do Energy Efficiency Investments Deliver at the Right Time?" National Bureau of Economic Research Working Paper No. 23097.

Borenstein, Severin. 2002. "The Trouble With Electricity Markets: Understanding California's Restructuring Disaster." *Journal of Economic Perspectives* 16 (1): 191–211.

Borenstein, Severin, and James Bushnell. 2018. "Do two electricity pricing wrongs make a right? Cost recovery, externalities, and efficiency." *Working paper.*

California Energy Commission. 2017. *Proposition 39: California Clean Energy Jobs Act, K-12 Program and Energy Conservation Assistance Act 2015-2016 Progress Report.* Technical report.

Cicala, Steve. 2015. "When does regulation distort costs? Lessons from fuel procurement in US electricity generation." *American Economic Review* 105 (1): 411–444.

———. 2017. "Imperfect Markets versus Imperfect Regulation in U.S. Electricity Generation." National Bureau of Economic Research Working Paper No. 23053.

Cooper, Adam. 2016. *Electric Company Smart Meter Deployments: Foundation for a Smart Grid.* Technical report. Institute for Electric Innovation.

Davis, Lucas, Alan Fuchs, and Paul Gertler. 2014. "Cash for coolers: evaluating a large-scale appliance replacement program in Mexico." *American Economic Journal: Economic Policy* 6 (4): 207–238.

Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif. 2019. *The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction.* Working Paper.

Eichholtz, Piet, Nils Kok, and John M. Quigley. 2013. "The Economics of Green Building." *Review of Economics and Statistics* 95 (1): 50–63.

Energy Information Administration. 2015. *Electric Power Monthly.* Technical report.

Engstrom, Ryan, Jonathan Hersh, and David Newhouse. 2016. "Poverty in HD: What Does High Resolution Satellite Imagery Reveal about Economic Welfare?" *Working Paper.*

Ferraro, Paul J., and Juan Jose Miranda. 2017. "Panel Data Designs and Estimators as Substitutes for Randomized Controlled Trials in the Evaluation of Public Programs." *Journal of the Association of Environmental and Resource Economists* 4 (1): 281–317.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. 2018. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program." *Quarterly Journal of Economics* 133 (3): 1597–1644.

Garg, Teevrat, Maulik Jagnani, and Vis Taraz. 2018. *Temperature and Human Capital in India.* Working Paper. UCSD.

Gerarden, Todd D, Richard G Newell, and Robert N Stavins. 2015. *Assessing the Energy-Efficiency Gap.* Technical report. Harvard Environmental Economics Program.

Gillingham, Kenneth, and Karen Palmer. 2014. "Bridging the energy efficiency gap: policy insights from economic theory and empirical evidence." *Review of Environmental Economics and Policy* 8 (1): 18–38.

Glaeser, Edward, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy." *American Economic Review: Papers & Proceedings* 106 (5): 114–118.

Graff Zivin, Joshua, Solomon M. Hsiang, and Matthew Neidell. 2017. "Temperature and Human Capital in the Short and Long Run." *Journal of the Association of Environmental and Resource Economists* 5 (1): 77–105.

Granderson, Jessica, Samir Touzani, Samuel Fernandes, and Cody Taylor. 2017. "Application of automated measurement and verification to utility energy efficiency program data." *Energy and Buildings* 142:191–199.

International Energy Agency. 2015. *World Energy Outlook.* Technical report.

Itron. 2017a. *2015 Custom Impact Evaluation Industrial, Agricultural, and Large Commercial: Final Report.* Technical report.

———. 2017b. *2015 Nonresidential ESPI Deemed Lighting Impact Evaluation: Final Report.* Technical report.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353:790–794.

Joskow, Paul L, and Donald B Marron. 1992. "What does a negawatt really cost? Evidence from utility conservation programs." *The Energy Journal* 13 (4): 41–74.

Kahn, Matthew, Nils Kok, and John Quigley. 2014. "Carbon emissions from the commercial building sector: The role of climate, quality, and incentives." *Journal of Public Economics* 113:1–12.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2017. "Human Decisions and Machine Predictions." *Working Paper.*

Kok, Nils, and Maarten Jennen. 2012. "The impact of energy labels and accessibility on office rents." *Energy Policy* 46 (C): 489–497.

Kotchen, Matthew J. 2017. "Longer-Run Evidence on Whether Building Energy Codes Reduce Residential Energy Consumption." *Journal of the Association of Environmental and Resource Economists* 4 (1): 135–153.

Kushler, Martin. 2015. "Residential energy efficiency works: Don't make a mountain out of the E2e molehill." *American Council for an Energy-Efficient Economy Blog.*

Levinson, Arik. 2016a. "How Much Energy Do Building Energy Codes Save? Evidence from California Houses." *American Economic Review* 106 (10): 2867–2894.

Levinson, Arik. 2016b. "How Much Energy Do Building Energy Codes Save? Evidence from California Houses." *American Economic Review* 106 (10): 2867–94.

McCaffrey, Daniel, Greg Ridgeway, and Andrew Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *RAND Journal of Economics* 9 (4): 403–425.

McKinsey & Company. 2009. *Unlocking energy efficiency in the U.S. economy.* Technical report. McKinsey Global Energy and Materials.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Myers, Erica. 2015. "Asymmetric information in residential rental markets: implications for the energy efficiency gap." *Working Paper.*

Naik, Nikhil, Ramesh Raskar, and Cesar Hidalgo. 2015. "Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance." *American Economic Review: Papers & Proceedings* 106 (5): 128–132.

Novan, Kevin, and Aaron Smith. 2018. "The Incentive to Overinvest in Energy Efficiency: Evidence from Hourly Smart-Meter Data." *Journal of the Association of Environmental and Resource Economists* 5 (3): 577–605.

Park, R. Jisung. 2017. *Hot Temperature and High Stakes Cognitive Assessments.* Working Paper. UCLA.

Ryan, Nicholas. 2018. "Energy Productivity and Energy Demand: Experimental Evidence from Indian Manufacturing Plants." National Bureau of Economic Research Working Paper No. 24619.

Varian, Hal R. 2016. "Causal inference in economics and marketing." *Proceedings of the National Academy of Sciences* 113 (27): 7310–7315.

Wyss, Richard, Alan Ellis, Alan Brookhart, Cynthia Girman, Michele Funk, Robert LoCasale, and Til Sturmer. 2014. "The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score." *American Journal of Epidemiology* 180 (6): 645–655.

Table 1: Average characteristics of schools in the sample

Characteristic	Untreated	Treated	T-U
Hourly energy use (kWh)	33.1	57.5	24.4
	(34.4)	(73.0)	[<0.01]
First year in sample	2012	2010	-2
	(1.7)	(1.8)	[<0.01]
Total enrollment	538	730	192
	(363)	(488)	[<0.01]
Acad. perf. index (200-1000)	789	794	6
	(99)	(89)	[0.21]
Bond passed, last 2 yrs (0/1)	0.3	0.2	-0.0
	(0.4)	(0.4)	[0.33]
Bond passed, last 5 yrs (0/1)	0.4	0.4	0.0
	(0.5)	(0.5)	[0.91]
High school graduates (%)	23.4	23.3	-0.1
	(12.2)	(11.6)	[0.89]
College graduates (%)	20.0	20.3	0.3
	(12.3)	(12.0)	[0.55]
Single mothers (%)	20.5	19.3	-1.3
	(19.1)	(18.4)	[0.17]
African American (%)	5.6	6.1	0.5
	(9.3)	(8.0)	[0.25]
Asian (%)	8.8	11.7	2.9
	(12.9)	(16.2)	[<0.01]
Hispanic (%)	42.4	43.4	1.1
	(28.8)	(26.8)	[0.41]
White (%)	34.7	30.8	-3.9
	(27.0)	(24.4)	[<0.01]
Average temp. (° F)	60.1	60.8	0.7
	(4.1)	(3.5)	[<0.01]
Latitude	37.7	37.5	-0.2
	(1.2)	(1.0)	[<0.01]
Longitude	-121.6	-121.2	0.4
	(1.0)	(1.1)	[<0.01]
Number of schools	958	912	

Notes: This table displays average characteristics of the treated and untreated schools in our sample. Standard deviations are in parentheses, with p-values of the difference between treated and untreated schools in brackets. "Untreated" schools underwent no energy efficiency upgrades for the duration of our sample. The "T-U" column compares treated schools to the schools that installed zero upgrades. Each row is a separate calculation, and is not conditional on the other variables reported here. There is substantial evidence of selection into treatment: treated schools tend to consume more electricity; have been in the sample longer; are larger; are in hotter locations; and are located southeast of untreated schools.

Table 2: Panel fixed effects results

	(1)	(2)	(3)	(4)	(5)	(6)
Treat × post	-2.90	-3.50	-2.23	-1.30	-1.81	-1.75
	(0.45)	(0.45)	(0.48)	(0.47)	(0.49)	(0.47)
Observations	55,818,652	55,817,256	55,817,256	55,818,652	55,817,256	55,821,180
Realization rate	0.68	0.81	0.90	0.54	0.75	0.74
	(0.11)	(0.10)	(0.20)	(0.19)	(0.20)	(0.20)
School-Hour FE	Yes	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes	Yes
Time trend	No	No	Yes	No	No	No
Month of Sample FE	No	No	No	Yes	Yes	Yes
Temp Ctrl	No	No	No	No	No	Yes

Notes: This table reports results from estimating Equation (3.1), with hourly energy consumption in kWh as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors, clustered at the school level, are in parentheses. Realization rates are calculated by dividing the regression results on a complementary regression of *ex ante* engineering energy savings where expected (and zero otherwise) on our treatment variable, where we include the same set of controls and fixed effects.

Table 3: Sensitivity of panel fixed effects results to outliers

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Trim outlier observation						
Realization rate	0.47	0.59	0.42	0.11	0.28	0.53
	(0.09)	(0.09)	(0.14)	(0.14)	(0.14)	(0.13)
Point estimate	-1.96	-2.49	-1.10	-0.28	-0.72	-1.38
	(0.37)	(0.37)	(0.36)	(0.36)	(0.36)	(0.34)
Observations	54,701,384	54,699,856	54,699,856	54,701,384	54,699,856	54,703,796
Panel B: Trim outlier schools						
Realization rate	0.71	0.85	0.81	0.44	0.64	0.67
	(0.11)	(0.11)	(0.18)	(0.18)	(0.18)	(0.18)
Point estimate	-2.70	-3.27	-1.91	-1.02	-1.49	-1.53
	(0.42)	(0.42)	(0.42)	(0.42)	(0.43)	(0.41)
Observations	55,058,188	55,056,816	55,056,816	55,058,188	55,056,816	55,060,740
Panel C: Trim observations and schools						
Realization rate	0.50	0.63	0.44	0.11	0.29	0.54
	(0.10)	(0.10)	(0.15)	(0.15)	(0.15)	(0.14)
Point estimate	-1.91	-2.43	-1.07	-0.26	-0.70	-1.31
	(0.38)	(0.38)	(0.36)	(0.36)	(0.37)	(0.34)
Observations	54,037,088	54,035,584	54,035,584	54,037,088	54,035,584	54,040,528
School-Hour FE	Yes	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes	Yes
Time trend	No	No	Yes	No	No	No
Month of Sample FE	No	No	No	Yes	Yes	Yes
Temp. Ctrl	No	No	No	No	No	Yes

Notes: This table reports results from estimating Equation (3.1), with hourly energy consumption in kWh as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors, clustered at the school level, are in parentheses. Realization rates are calculated by dividing the regression results on a complementary regression of ex-ante engineering energy savings where expected (and zero otherwise) on our treatment variable, also including the same set of controls. In Panel A, we drop observations below the 1st or above the 99th percentile of the dependent variable: energy consumption. In Panel B, we drop schools below the 1st or above the 99th percentile of expected savings. In Panel C, we drop both.

Table 4: Machine learning results

	(1)	(2)	(3)	(4)	(5)
Treat × post	-3.86	-4.17	-3.43	-2.24	-2.61
	(0.51)	(0.53)	(0.50)	(0.48)	(0.50)
Observations	55,822,576	55,821,180	55,821,180	55,822,576	55,821,180
Realization rate	0.90	0.96	1.01	0.70	0.78
	(0.12)	(0.12)	(0.15)	(0.15)	(0.15)
School-Hour FE	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes
Time trend	No	No	Yes	No	No
Month of Sample FE	No	No	No	Yes	Yes

Notes: This table reports results from estimating Equation (3.2), with prediction errors in hourly energy consumption in kWh as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors, clustered at the school level, are in parentheses. Realization rates are calculated by dividing the regression results on a complementary regression of ex-ante engineering energy savings where expected (and zero otherwise) on our treatment variable, also including the same set of controls. All regressions include a control for being in the post-training period for the machine learning.

Table 5: Sensitivity of machine learning results to outliers

	(1)	(2)	(3)	(4)	(5)
Panel A: Trim outlier observation					
Realization rate	0.84	0.89	0.87	0.65	0.70
	(0.08)	(0.09)	(0.09)	(0.09)	(0.09)
Point estimate	-3.47	-3.68	-3.05	-2.20	-2.45
	(0.34)	(0.35)	(0.32)	(0.31)	(0.32)
Observations	54,706,124	54,704,668	54,704,668	54,706,124	54,704,668
Panel B: Trim outlier schools					
Realization rate	0.94	1.01	1.02	0.66	0.76
	(0.13)	(0.13)	(0.15)	(0.15)	(0.16)
Point estimate	-3.62	-3.93	-3.16	-1.98	-2.35
	(0.48)	(0.51)	(0.48)	(0.45)	(0.48)
Observations	55,062,112	55,060,740	55,060,740	55,062,112	55,060,740
Panel C: Trim observations and schools					
Realization rate	0.89	0.94	0.91	0.67	0.72
	(0.09)	(0.09)	(0.10)	(0.10)	(0.10)
Point estimate	-3.35	-3.55	-2.94	-2.10	-2.35
	(0.34)	(0.35)	(0.32)	(0.31)	(0.32)
Observations	54,020,156	54,018,720	54,018,720	54,020,156	54,018,720
School-Hour FE	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes
Time trend	No	No	Yes	No	No
Month of Sample FE	No	No	No	Yes	Yes

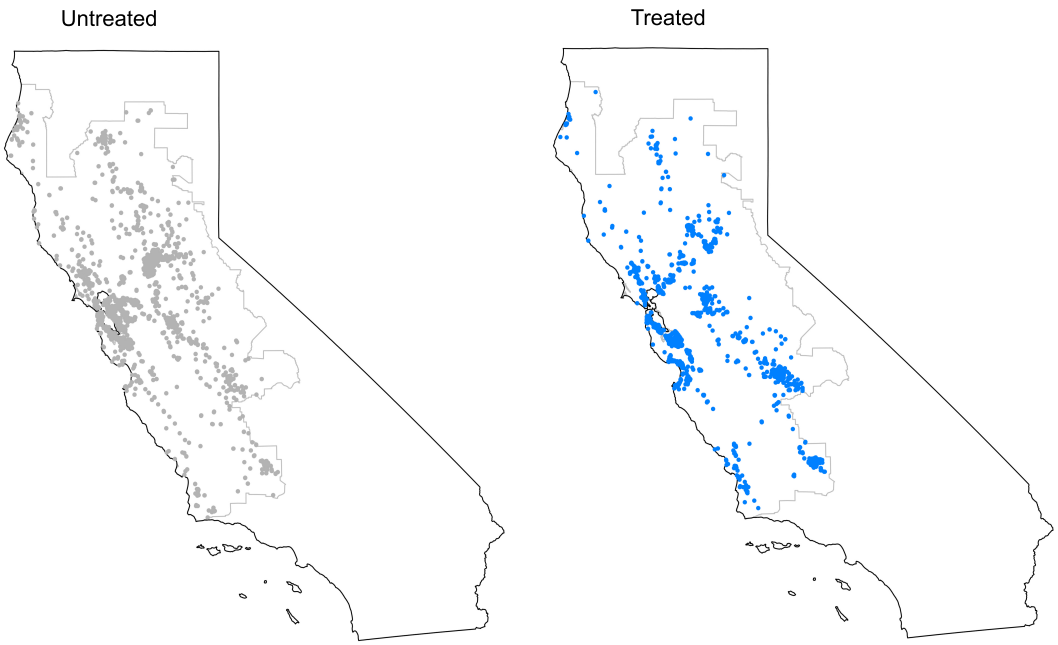
Notes: This table reports results from estimating Equation (3.2), with prediction errors in hourly energy consumption in kWh as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors, clustered at the school level, are in parentheses. Realization rates are calculated by dividing the regression results on a complementary regression of ex-ante engineering energy savings where expected (and zero otherwise) on our treatment variable, also including the same set of controls. All regressions include a control for being in the post-training period for the machine learning. In Panel A, we drop observations below the 1st or above the 99th percentile of the dependent variable: energy consumption. In Panel B, we drop schools below the 1st or above the 99th percentile of expected savings. In Panel C, we drop both.

Table 6: Predicting heterogeneous effects

Variable	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.78	0.36	0.52	0.59	0.57	0.55
	(0.12)	(0.28)	(0.22)	(0.24)	(0.35)	(0.30)
HVAC only (0/1)		0.70	0.69	0.51	0.76	0.86
		(0.34)	(0.27)	(0.29)	(0.42)	(0.37)
Lighting only (0/1)		0.70	0.50	0.41	0.47	0.60
		(0.37)	(0.29)	(0.32)	(0.46)	(0.39)
HVAC and Lighting (0/1)		0.37	0.19	0.10	0.02	0.40
		(0.35)	(0.28)	(0.31)	(0.45)	(0.39)
Longitude			-0.09	-0.10	-0.26	-0.25
			(0.21)	(0.23)	(0.33)	(0.29)
Latitude			0.28	0.30	0.23	0.32
			(0.15)	(0.16)	(0.24)	(0.21)
Average temperature (° F)			0.15	0.16	0.30	0.33
			(0.18)	(0.20)	(0.28)	(0.24)
Total enrollment				0.34	0.23	0.32
				(0.09)	(0.12)	(0.11)
Academic perf. index (200-1000)					-0.23	-0.24
					(0.16)	(0.14)
Poverty rate					-0.15	-0.10
					(0.16)	(0.14)
Expected savings (kWh)						-0.20
						(0.09)
Number of schools	903	903	881	838	776	776

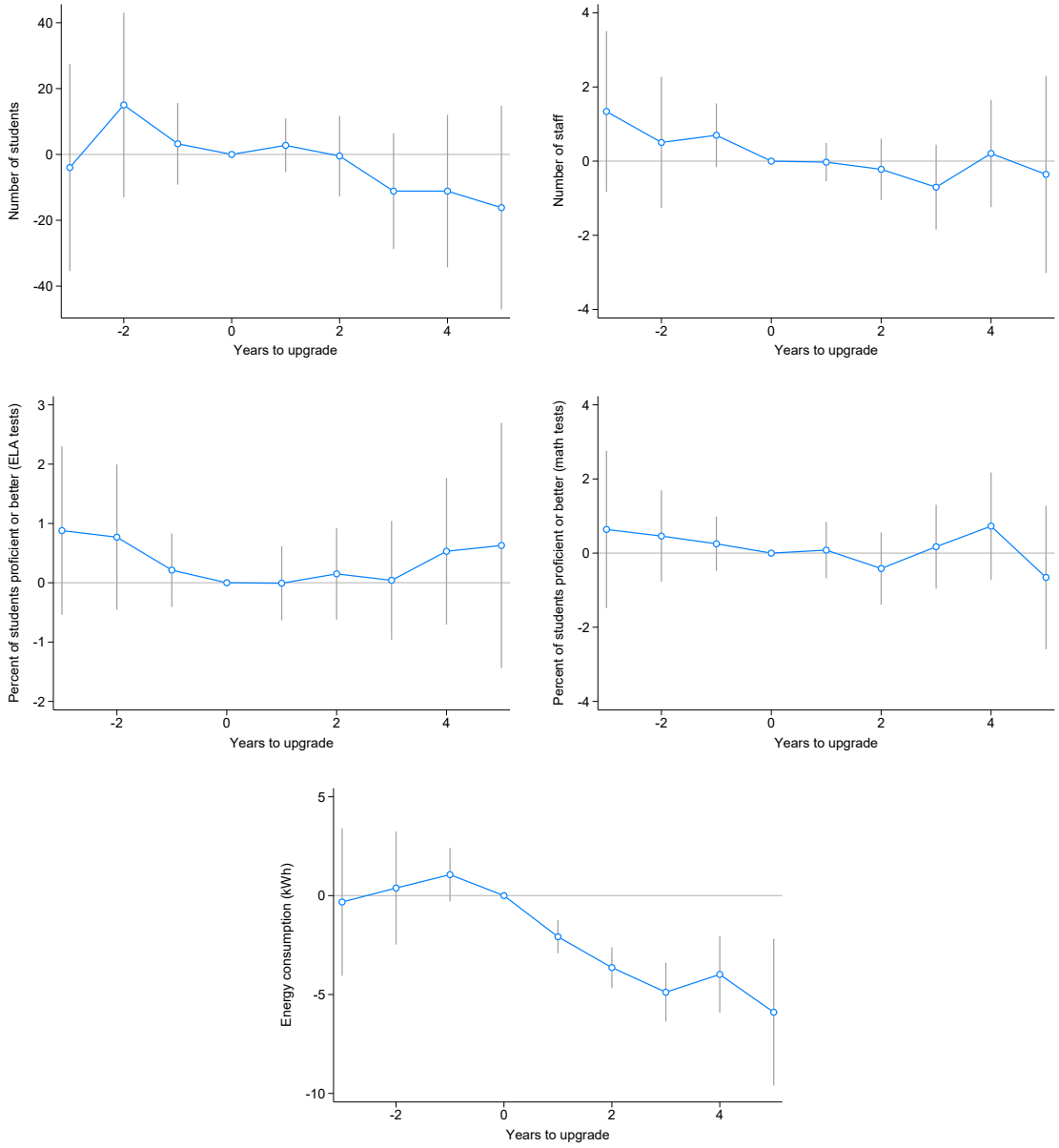
Notes: This table presents results from median regressions of school-specific realization rates on a variety of covariates. The school-specific realization rates are estimated from a regression of prediction errors (in kWh) on school-specific treatment indicators and school-by-hour-by-month fixed effects. This table presents results for treated schools only. All estimates are weighted by the number of observations at each school. All variables (except dummy variables) are normalized to have mean zero and a standard deviation of one. Standard errors, robust to heteroskedasticity, are in parentheses.

Figure 1: Locations of untreated and treated schools



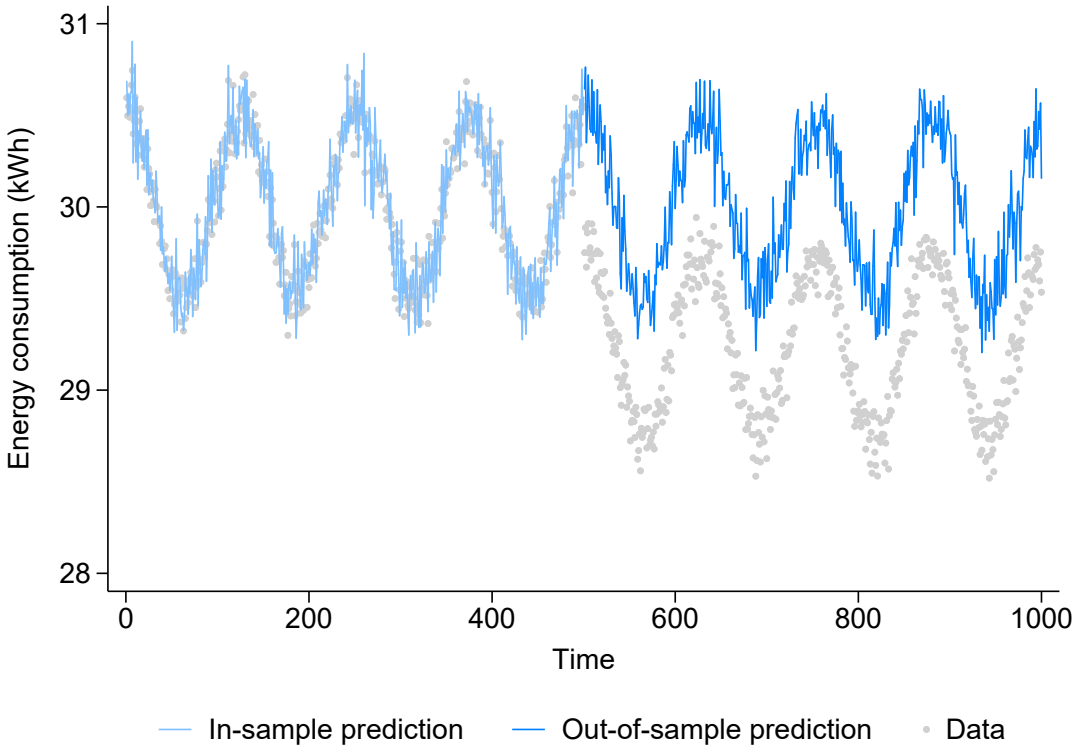
Notes: This figure displays the locations of schools in our sample. "Untreated" schools, in gray on the left, did not undertake any energy efficiency upgrades during our sample period. "Treated" schools, in blue on the right, installed at least one upgrade during our sample. There is substantial overlap in the locations of treated and untreated schools. The light gray outline shows the PG&E service territory.

Figure 2: School characteristics before and after treatment



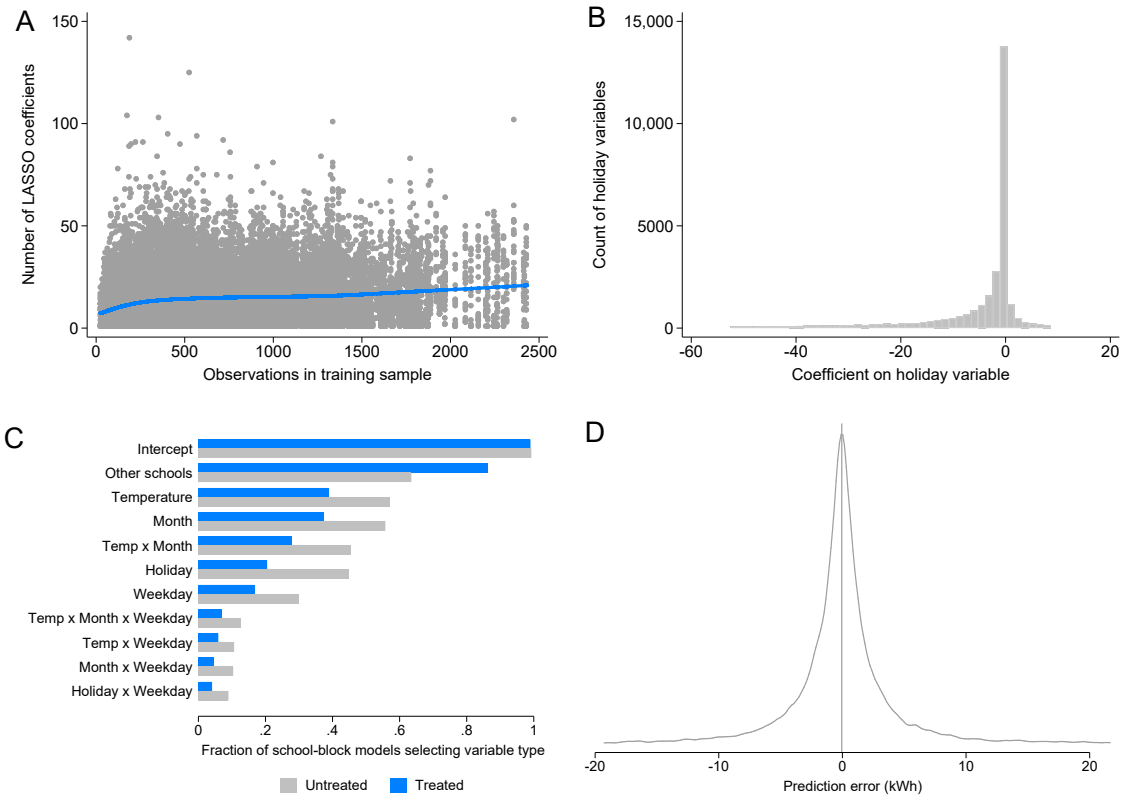
Notes: This figure shows point estimates and 95 percent confidence intervals from event study regressions of school demographics and test scores before and after an energy efficiency upgrade. We normalize time relative to the year each school undertook its first upgrade. Standard errors are clustered by school. The top left panel displays results for number of students enrolled in school; the top right panel shows results for number of staff members; the middle left panel shows results for the percent of students scoring proficient (the state standard) or better on California's Standardized Testing And Reporting (STAR) math tests; and the middle right panel shows results for the percent of students scoring proficient or better on California's STAR English and Language Arts (ELA) tests. We find no strong evidence of substantial changes in any of these variables around the timing of energy efficiency upgrades. We find evidence that upgrades effectively reduced energy consumption at treated schools, as shown in the bottom panel.

Figure 3: Machine learning approach: graphical intuition



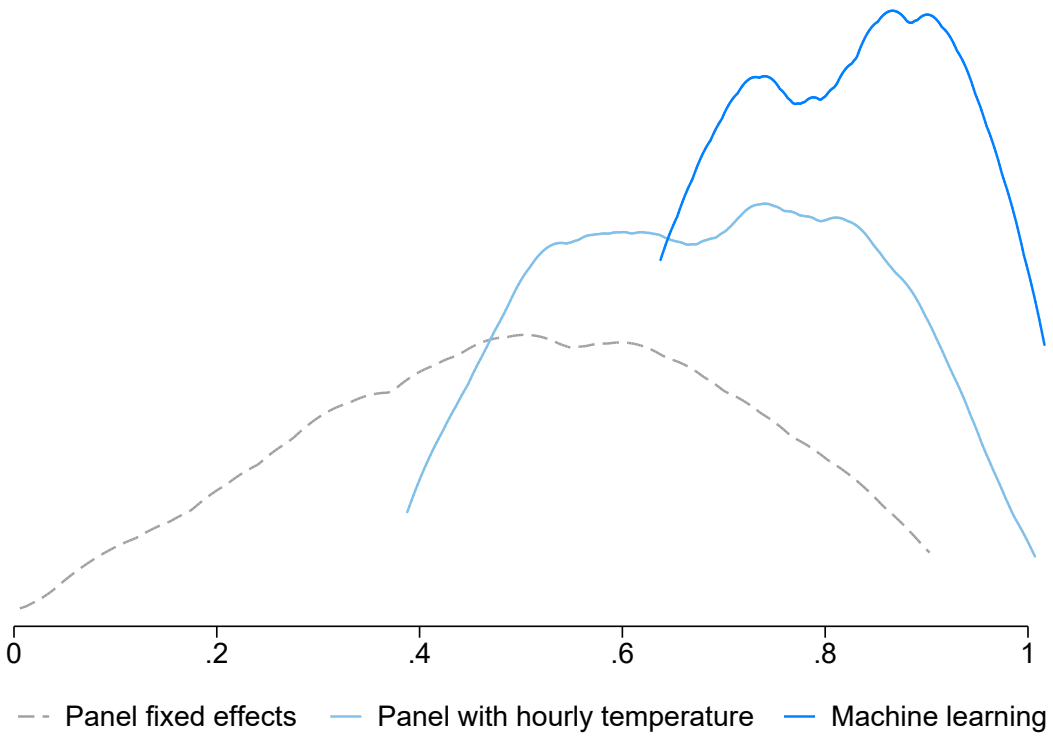
Notes: This figure displays a stylized overview of how our machine learning approach works. In step 1, we use the pre-treatment data only to fit a school-specific machine learning model of energy consumption (light blue line). We then use these models model to create fully out-of-sample predictions of counterfactual energy use in the post-treatment period (dark blue line). We compare the post-treatment counterfactuals to the actual data (gray points) to compute prediction errors. If the method is performing properly, these prediction errors will be close to zero in the untreated group. Non-zero prediction errors in the treatment group correspond to treatment effects. In step 2, we use these prediction errors as the dependent variable in a panel fixed effects regression.

Figure 4: Machine learning diagnostics



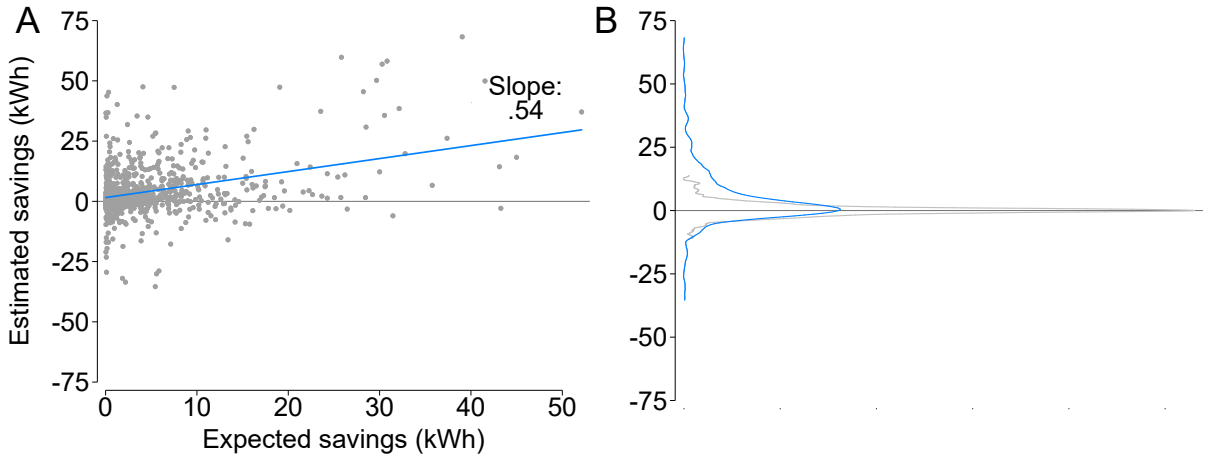
Notes: This figure presents three checks of our machine learning methodology. Panel A displays the relationship between the number of observations in the pre-treatment ("training") dataset and the number of variables LASSO selects to include in the prediction model for each school in the sample. Schools with very few training observations yield sparse models. As expected, the larger the training sample, the more flexible the prediction model becomes up to a point. This suggests that the LASSO is not overfitting, but that the underlying data generating process is relatively sparse, which is required for the LASSO to perform well. Panel B displays the marginal effect of holiday indicators in each school-specific prediction model. The majority of the coefficients on these models are negative and we do not observe large outliers, which suggests that the LASSO model is picking up patterns that we would expect to be present in the data and that will do well out of sample. Panel C displays the categories of variables selected by our preferred LASSO method for untreated and treated schools. Most models selected at least one untreated school's prediction for inclusion in the model. Finally, Panel D shows the distribution of average prediction errors out-of-sample for untreated schools (trimming the top and bottom 1 percent), which are centered around zero.

Figure 5: Comparison of methods across specifications and samples



Notes: This figure shows the distribution of implied realization rates using three alternative approaches: a panel fixed effects regression, a panel fixed effects regressions with school-specific temperature controls, and a machine learning approach. For each of these three approaches, we consider five specifications (the ones in the main Tables (3.1) and (3.2). Each of the five specifications is estimated on five different samples: no trimming; trimming observations below the 1st (2nd) and above the 99th (98th) percentile of the dependent variable; trimming the schools with smallest and largest 1 percent of interventions; and a combination of the latter two 1 percent trims. Each kernel density is computed from a total of 25 estimates.

Figure 6: School-specific effects



Notes: This figure displays school-specific savings estimates. We generate these estimates by regressing prediction errors in kWh onto an intercept and school-by-post-training dummies. The coefficients on these dummies are the savings estimates. Panel A compares estimated savings with expected savings among treated schools only. This method produces a realization rate of 0.54 (weighted by the number of observations per school after removing outliers in expected savings), though there is substantial heterogeneity. Panel B displays kernel densities of estimated savings in the untreated group (gray line) and estimated savings in the treated group (blue line). While the distribution of effects in the untreated is narrow and centered around zero, the treated group appears shifted towards more savings.

Appendix: For online publication

A Supplemental tables and figures

Table A.1: Panel fixed effects results (alternative standard errors)

Clustering	(1)	(2)	(3)	(4)	(5)	(6)
	-2.90	-3.50	-2.23	-1.30	-1.81	-1.83
School	(0.45)	(0.45)	(0.48)	(0.47)	(0.49)	(0.48)
School, month of sample	[1.59]	[0.72]	[0.54]	[0.54]	[0.50]	[0.49]
Observations	55,818,652	55,817,256	55,817,256	55,818,652	55,817,256	55,821,180
School-Hour FE	Yes	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes	Yes
Time trend	No	No	Yes	No	No	No
Month of Sample FE	No	No	No	Yes	Yes	Yes
Temp. Ctrl	No	No	No	No	No	Yes

Notes: This table reports results from estimating Equation (3.1), with hourly energy consumption in kWh as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. This table shows two variations on clustered standard errors: errors clustered at the school level, as in the main text, in parentheses; and errors clustered at the school and month-of-sample level, in brackets.

Table A.2: Matching results

	(1)	(2)	(3)	(4)	(5)
Any district	-2.70	-2.99	-0.66	-0.31	-0.47
	(0.93)	(0.98)	(1.07)	(1.01)	(1.13)
Same district	-0.17	-0.40	1.14	0.97	0.92
	(0.83)	(0.82)	(0.86)	(0.79)	(0.84)
Opposite district	-3.55	-3.64	-0.44	-0.16	-0.03
	(0.94)	(1.01)	(1.12)	(1.05)	(1.18)
Observations	6,043,046	6,042,653	6,042,653	6,043,046	6,042,653
School-Hour FE	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes
Time trend	No	No	Yes	No	No
Month of Sample FE	No	No	No	Yes	Yes

Notes: This table reports results from estimating Equation (3.1), with hourly energy consumption in kWh as the dependent variable. As above, the independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. The untreated group in these regressions is chosen via nearest-neighbor matching. In particular, we match one untreated school to each treated school. Each row in the table employs a different restriction on which schools are allowed to be matched to any given treatment school. "Any district" matches allow any untreated school to be matched to a treatment school; "same district" matches are restricted to untreated schools in the same school district, and "opposite district" matches are restricted to untreated schools from different districts. In each case, the matching variables are the mean, maximum, and standard deviation of electricity consumption in three hour blocks (e.g., 9 AM-Noon) from the pre-treatment period; demographic variables measured at the census block level, including the poverty rate, log of per capita income, school-level variables (enrollment; age of the school; grades taught; an academic performance index; and climate). These estimates are relatively sensitive to which schools are included. Standard errors, clustered at the school level, are in parentheses.

Table A.3: R^2s of prediction models across machine learning methods

	(1)	(2)	(3)	(4)	(7)	(8)
10th percentile	-0.02	0.09	0.01	0.13	0.11	-1.85
25th percentile	0.24	0.24	0.31	0.32	0.28	-0.15
50th percentile	0.49	0.46	0.63	0.62	0.50	0.43
75th percentile	0.69	0.65	0.86	0.85	0.66	0.64
90th percentile	0.81	0.78	0.95	0.94	0.77	0.74
Method	LASSO	LASSO	LASSO	LASSO	RF	RF
Basic variables	X	X	X	X	X	X
Hour-specific model	X	X	X	X	X	
Untreated schools $-i$			X	X		
Tuning parameter	Min	1SE	Min	1SE		

Notes: This table reports the R^2 of the prediction models for untreated schools during the post-treatment period. As that these predictions are completely out-of-sample, and therefore extreme outliers may be a concern, we present the distribution of the R^2. Columns 1 through 6 display predictions generated via LASSO, while Columns 7 and 8 show predictions generated using a random forest algorithm. In all but Column 8, we generate prediction models for each school-hour separately. The "basic variables" include day of the week, a holiday dummy, a seasonal spline, a temperature spline, and all of their their multi-way interactions. In Columns 3, 4, 5, and 6, we include energy consumption at all (other) untreated schools as candidate variables. For the LASSO estimates, we report results for two tuning parameters: "Min," which minimizes the root mean squared error, or "1SE," which chooses a slightly more parsimonious model than Min, but which has a root mean squared error that remains within one standard error of Min. Overall, we find that the LASSO model where we allow for both basic variables and untreated school consumption, with a 1SE tuning parameter, provides the best overall fit. Note that some of the R^2s are negative. This is not surprising, given that these are fully-out-of-sample predictions.

Table A.4: Machine learning results (alternative standard errors)

Clustering	(1)	(2)	(3)	(4)	(5)
	-3.86	-4.17	-3.43	-2.24	-2.61
School	(0.51)	(0.53)	(0.50)	(0.48)	(0.50)
School, month of sample	[0.67]	[0.63]	[0.54]	[0.49]	[0.51]
Observations	55,822,576	55,821,180	55,821,180	55,822,576	55,821,180
School-Hour FE	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes
Time trend	No	No	Yes	No	No
Month of Sample FE	No	No	No	Yes	Yes

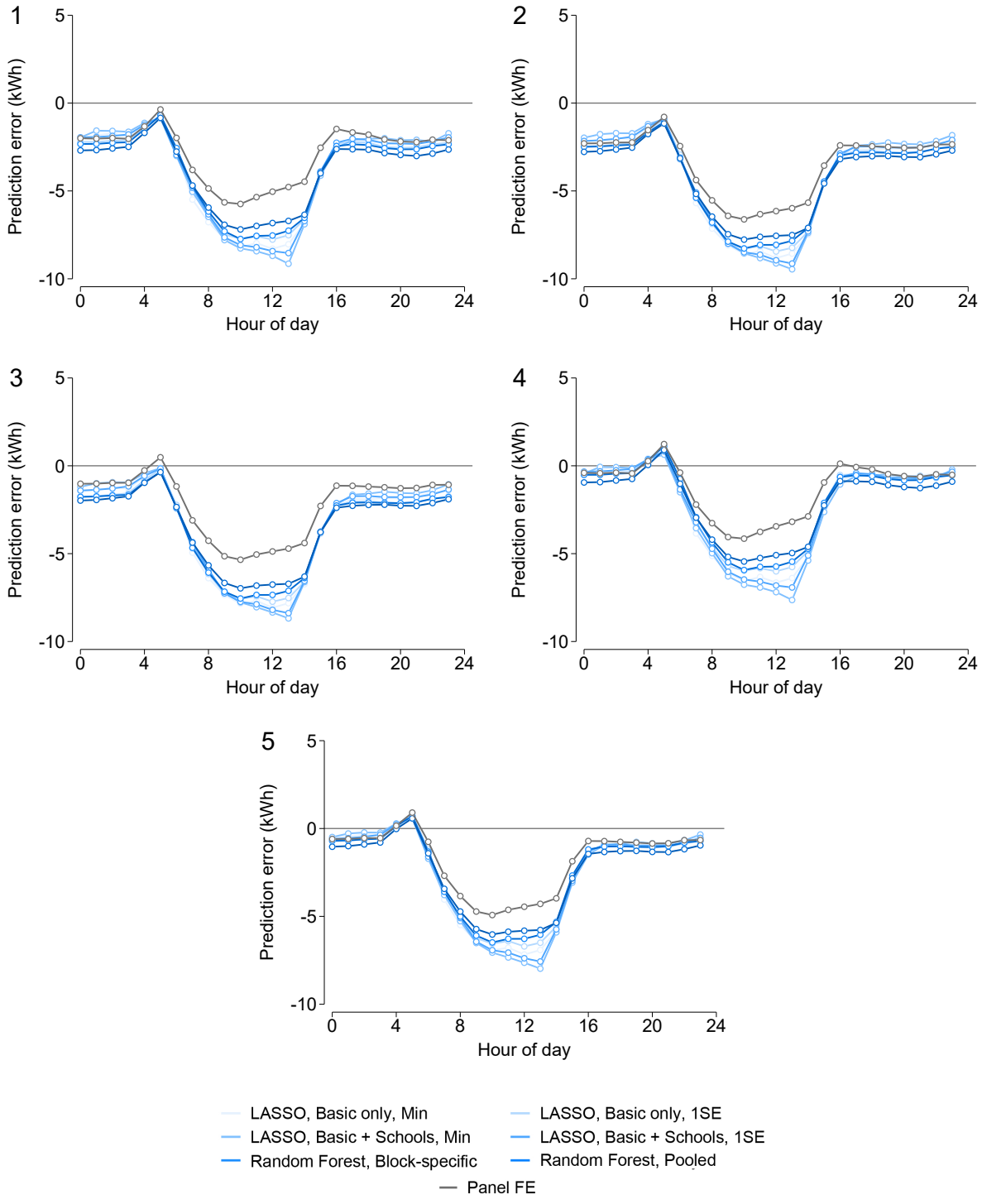
Notes: This table reports results from estimating Equation (3.2), with prediction errors in hourly energy consumption in kWh as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. This table shows two variations on clustered standard errors: errors clustered at the school level, as in the main text, in parentheses; and errors clustered at the school and month-of-sample level, in brackets. All regressions include a control for being in the post-training period for the machine learning.

Table A.5: Machine learning results (alternative prediction methods)

	(1)	(2)	(3)	(4)	(5)	(6)
Treat × post	-2.57	-2.51	-2.64	-2.61	-2.44	-2.51
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.51)
Realization rate	0.77	0.76	0.79	0.78	0.73	0.75
	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)
Method	LASSO	LASSO	LASSO	LASSO	RF	RF
Hour-specific model	X	X	X	X	X	
Basic variables	X	X	X	X	X	X
Untreated schools $-i$			X	X		
Tuning parameter	Min	1SE	Min	1SE		

Notes: This table reports results from estimating Equation (3.2), with prediction errors in hourly energy consumption in kWh as the dependent variable. All regressions include school-by-hour and month-of-sample fixed effects. Each column displays results from a different prediction approach. Columns 1 through 6 display predictions generated via LASSO, while Columns 7 and 8 show predictions generated using a random forest algorithm. In all but Column 8, we generate prediction models for each school-hour separately. The "basic variables" include day of the week, a holiday dummy, a seasonal spline, a temperature spline, and all of their their multi-way interactions. In Columns 3, 4, 5, and 6, we include energy consumption at all (other) untreated schools as candidate variables. For the LASSO estimates, we report results for two tuning parameters: "Min," which minimizes the root mean squared error, or "1SE," which chooses a slightly more parsimonious model than Min, but which has a root mean squared error that remains within one standard error of Min. In all cases, the independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. All regressions include a control for being in the post-training period for the machine learning, and standard errors are clustered at the school level.

Figure A.1: Machine learning results by hour (alternative prediction methods)



Notes: This figure presents treatment effects for each hour of the day estimated using prediction errors based on electricity consumption in kWh as the dependent variable. Here, we present results from 9 different estimation procedures: LASSOs with, without, and exclusively using other schools' consumption as candidate variables using a larger and smaller tuning parameter; random forests with and without imposing hour-specific branches; and the panel fixed effects analogue. Each panel corresponds to one column of Table 4.