Using Hurdle Models for Long-term Projections of Residential Solar Photovoltaic Systems Installations

Janice Lent
U.S. Energy Information Administration,[1]
1000 Independence Ave, S.W., Washington, DC 20585

Abstract
We describe research on the use of hurdle models for projecting the number of installations of residential solar photovoltaic (PV) systems in the United States. The U.S. Energy Information Administration (EIA) publishes detailed energy-related projections annually in its *Annual Energy Outlook* (*AEO*). The 2017 edition of the *AEO* provides projections to 2050. The EIA uses its own National Energy Modeling System (NEMS) to produce *AEO* projections.

Hurdle models have been used to model count data in various settings, including public health and econometric applications. Rothfield (2010) used hurdle models to identify significant drivers of residential PV installations in California, revealing both economic and social effects. We use the GAMLSS package in R to fit hurdle models, incorporating logistic and negative binomial components, to zipcode-level residential PV installation data. We combine the model coefficients with projected variables from the NEMS to project future PV installations. The projections are aggregated to the national and Census Division levels for publication.

Key Words: Hurdle model, negative binomial regression, logistic regression, GAMLSS

1. The Rising Penetration of Residential Solar Photovoltaic Systems

Solar photovoltaic (PV) systems are the most widely used solar electricity generating systems both globally and in the United States. In 2014, solar PV accounted for 6.9% of net electricity consumption in Germany and met 2.9% of Japan's electricity demand. PV systems range in size from small residential systems (10 killowatts or less) to utility-scale systems generating more than one megawatt. Unlike solar thermal systems, which collect solar heat and use it to power conventional generators, PV systems generate electric current from sunlight and can therefore operate at cool temperatures without losing efficiency with reduced scale. Cloud cover and other sources of shade, however, affect the efficiency of solar PV systems.

Because of declining system costs, improving technologies, and government policies, PV penetration has increased substantially in the United States, where small-scale PV electricity generation more than doubled between 2014 and 2016. PV systems now account for over 90% of installed U.S. solar electricity generating capacity. In 2014, about 75% of

[1] Disclaimer: Opinions expressed in this paper are those of the author and do not constitute policy of the U.S. Energy Information Administration.

installed solar PV electric generating capacity was concentrated in five states,[2] but installations outside these states are increasing. Stanford University futurist Dr. Tony Seba has predicted that, by 2030, nearly all American homes and businesses will have solar PV systems (Seba 2014).

Both state and federal tax incentives have made solar PV more economically competitive in recent years. The federal investment tax credit (ITC), implemented in 2006, effectively provides a 30% rebate on the cost of purchasing and installing a residential PV system. Many states also have renewable portfolio standards (RPS) that offer solar renewable energy certificates (SREC) for electricity generated through solar PV. The SREC's may be sold to electricity suppliers, who are required to hold a minimum number of them in order to comply with the RPS.

Solar technology, including PV technology, is less cost effective in the residential sector than in the utility sector. The cost of an installed PV system is the sum of the cost of the actual PV panels and the balance of system (BOS) costs. Because of hardware and logistical costs, coupled with relatively weak competition among residential PV installers, the BOS component can make up more than two thirds of the total system cost for residential installations.[3] The high upfront cost has led to power purchase agreements (PPA) and leasing arrangements through which a third party developer installs and retains ownership of a PV system on a home, and the homeowner either pays a monthly lease amount or buys the installed PV system's electricity from the developer.

Under a PPA, the price that the homeowner pays for the electricity is based on the retail electricity rate for the location and a discount factor which generally saves the homeowner approximately 15% on the electricity generated by the PV system rather than purchased from the grid. PPA and leasing arrangements minimize or remove the upfront cost of a PV system to homeowners, while tax and RPS incentives help make the investment worthwhile for the third party developers. Although the trend toward PPA arrangements appears to be weakening, the majority of new residential PV systems installed in 2013 and 2014 were covered by third party agreements.

The growth in the PV market, along with the implementation of government incentives, has led to increased public-access data sources on PV prices and installations. Several states now publish project-level administrative records, and the National Renewable Energy Laboratory (NREL) has launched the Open PV Project, which provides real time data on the status of PV penetration nationwide (https://openpv.nrel.gov/). The modeling approach described here relies on data from multiple sources with the limitations noted below. EIA expects to further update and develop the models in the future as the available data sources expand.

[2] Key Figures of the "Solar Market Insight Report 2015 Q2," Solar Energy Industry Association, available at http://www.seia.org/research-resources/solar-market-insight-report-2015-q2.
[3] MIT Energy Initiative report, "The Future of Solar Energy: An Interdisciplinary MIT Study led by the MIT Energy Initiative," available at https://mitei.mit.edu/futureofsolar.

2. Modeling Residential Solar PV in the National Energy Modeling System

EIA's National Energy Modeling System (NEMS) is a modular system incorporating energy supply, consumption, and integrating modules. Prior to 2017, the NEMS Residential Energy Consumption Module (RECM) relied on data from EIA's 2009 Residential Energy Consumption Survey (RECS). The RECS provides data on PV installations, electricity prices, and housing unit characteristics such as square footage. The RECS data were mapped to an NREL database containing zipcode-level estimates of solar irradiation levels. Projections of new solar PV installations were based on a cash flow analysis that calculated the number of years required for a residential PV installation to "pay for itself" in the form of lower electricity costs. A logistic curve, with parameters assigned by expert judgment, was then used to project future installations based on the number of years to achieve payoff.

The new projection method, implemented in the 2017 version of the NEMS, combines zipcode level data from three state databases, NREL, the Census Bureau's American Community Survey (ACS), and other sources. Previous studies matching data from solar PV databases to ACS data include those by Rothfield (2010) and Hernandez (2013). Rothfield presents a statistical model-based investigation of the effects of economic and social factors on residential PV installation decisions in California. The results indicate that previous installations within a zipcode significantly increase the likelihood of future installations, even when the models control for other factors (e.g., income, education of householders). Through a descriptive analysis of the installation data for California, New Jersey, and Arizona, Hernandez shows the effects of income on homeowners' decisions to adopt solar PV.

3. Using Hurdle Models to Project Residential PV Installations

Online databases of residential solar PV installations are available for the states of California, New Jersey, and Arizona.[4] NREL publishes an Open PV Database, containing voluntarily-reported data on new PV installations. The Open PV data for Massachusetts and Maryland are sufficiently complete to be used in the models. These five states therefore serve as "data states." The five states combined represent a broad range of solar irradiation levels and economic conditions. We refer to the zipcodes within these states as *data zipcodes*. We use the combined data from these zipcodes, along with covariate data available for most zipcodes in the United States, to estimate historical PV installations for zipcodes outside the data states. We refer to the zipcodes for which we wish to estimate historical installations as *target zipcodes*. All data used in the hurdle models are annual.

The zipcode-level data are for Census Bureau zipcode tabulation areas. Although these include the vast majority of zipcodes, some zipcodes with few or no residential units are

---

[4] The residential PV installations data can be downloaded from the following sites: http://www.californiasolarstatistics.ca.gov/current_data_files/, http://www.njcleanenergy.com/renewable-energy/project-activity-reports/installation-summary-by-technology/solar-installation-projects, and www.arizonagoessolar.org. The three test states represent a broad range of solar insolation levels and median incomes. In the Census Bureau's 2012-2014 state income rankings, New Jersey had the 4th highest median household income among all states ($64,670), while California ranked 14th ($56,883), and Arizona ranked 33rd ($49,562).

not treated as tabulation areas in Census Bureau surveys. Also, the residential solar PV installation data may be subject to some under counting, as discussed in [1]. The estimated PV capacity and generation estimates are therefore benchmarked to EIA's historical estimates. The goal is to produce projections of residential solar PV installations at the Census Division level that reflect assumptions regarding changes in median household income, PV prices, and other economic factors, as well as the "social spillover" effect documented by Rothfield (2010). Zipcode-level American Community Survey (ACS) data were downloaded from the Census Bureau website using the American Fact Finder (AFF) extraction tool.

To initialize the data series for the zipcodes outside the five datastates, we matched each target zipcode with a similar data zipcode. We used the historical data from the matching data zipcode, adjusted by a ratio of fitted model values, to impute historical installations for the target zipcode. For all but four of these target zipcodes, we used the hurdle model described in Section 4 to impute and project the number of residential PV installations. Because of extreme covariate values, which may have created a "snowball effect," the lagged dependent variable was omitted in the model for four zipcodes (three in Hawaii and one in Texas). The main steps in the method may thus be summarized as follows:

1. Obtain a covariate vector for each zipcode. Match each target zipcode to a data zipcode such that the Euclidean distance between the covariate vectors of the target zipcode and its matched data zipcode is minimized.
2. Fit a reduced hurdle model, including all covariates listed in subsection 4.1 (but excluding a lagged dependent variable), to the combined zipcode-level data for the data states.
3. For each target zipcode, impute the installation and lagged installation values from the matching data zipcode, adjusted by the ratio of the fitted values from the reduced model and a size adjustment factor (number of households).
4. Use the coefficients of the hurdle model described in Section 4, along with the zipcode-level covariate data and the imputed installation data, to estimate installations for the target zipcodes. (In the case of the four zipcodes with extreme covariate values, omit the lagged dependent variable from the model.)
5. Use the coefficients of the hurdle models, along with projected covariate data from the NEMS, to project residential PV installations at the zipcode level; aggregate the projections to the national and Census Division levels.

The models described in Section 4 incorporate zipcode-level ACS estimates of median household income and numbers of households. They account for solar irradiation and electricity prices at the zipcode level and for national-level annual mortgage interest rates and solar PV prices, aggregated to compute an average monthly payment for a solar PV system. The monthly payment is assumed constant across areas but varies over time. Other ACS zipcode-level variables that were tested in the models include median age and educational attainment of householders, as well as percentage of owner-occupied housing units. These were found insignificant in the models, due to high correlation with median household income. EIA state-level estimates of household electricity consumption were also tested as a proxy for electricity demand but proved too strongly (negatively) correlated with electricity prices to be included in the models. Similarly, one-year lagged measures of state-level average cooling degree days, obtained from the National Oceanic and Atmospheric Administration, were tested as proxy demand measures but were found insignificant.

Because the models are intended to be used for long-term projections (step 5 above), the available covariates are limited to those projected in the NEMS. Solar PPA and leasing rates, for example, are not directly available but are reasonably assumed to be driven by retail electricity rates, prices of installed PV panels, and interest rates, which are projected in the NEMS. The models cannot directly account for state and local policy effects, because these may be confounded with other differences between geographic areas (e.g., differences in retail electricity prices and/or solar irradiation levels). Future research, discussed in Section 6, may focus on incorporating policy effects into the projections through adjustments applied to the hurdle model estimates.

The hurdle models include all the variables used in the previous NEMS "payback model" except residential roof area. Zipcode-level population density estimates are used as a proxy for roof area. Roof area is expected to become less of a limiting factor in installation decisions as the solar PV technology evolves. For example, solar skyscrapers are being built with PV panels mounted on exterior walls and in windows. Concentrated solar PV (CPV) technology, which requires less roof area and is currently used in utility and large-scale commercial applications, may eventually become economically viable for residential projects. Residential system capacity calculations, separate from the statistical models described here, will continue to be performed in the NEMS.

4. Model Specification

4.1 Model Explanatory Variables

The model covariates are based on data from EIA, the National Renewable Energy Laboratory (NREL), the Census Bureau's American Community Survey (ACS), and the decennial census. The following covariates were included for each year t and zipcode z:

$x_{1,t,z}$ = median household income, estimated from the ACS and decennial census data;[5]

$x_{2,z}$ = annual average solar irradiation level, in kilowatthours per square meter per day (estimated by NREL as described in [8]), assumed constant over time; [6]

$x_{3,t,z}$ = electricity rate (cents per kilowatthour), estimated as described in Appendix A;

$x_{4,t,z}$ = number of households, estimated from the ACS and decennial census;

[5] ACS data are available for 2011 through 2013. Decennial census data were used, along with the ACS data, to interpolate median income and numbers of households for 2001 through 2010.

[6] The zipcode-level solar irradiation levels, downloaded from the NREL website, assume PV panels at a lateral tilt. The NREL database excludes the state of Alaska, because cloud cover obscures the Alaska irradiation measures gathered from satellites. Alaska solar irradiation data were imputed at the zipcode level, based on NREL data available for 17 locations within the state.

$x_{5,t}$ = installed price of solar PV panels in year t;[7] and

$x_{6,t}$ = annual average mortgage interest rate for year t.[8]

We combined the solar PV panel price ($x_{5,t}$) with the annual average mortgage interest rate ($x_{6,t}$) to compute a *monthly payment* ($x_{7,t}$) per kilowatt of installed capacity, based on a 30-year (360 month) mortgage:

$$x_{7,t} = 1{,}000 x_{5,t} \left[\frac{\frac{x_{6,t}}{12}\left(1 + \frac{x_{6,t}}{12}\right)^{360}}{\left(1 + \frac{x_{6,t}}{12}\right)^{360} - 1} \right].$$

We further combined the estimated number of households $x_{4,t,z}$ with land area information to create a measure of population density in units of households per square mile:

$$x_{8,t,z} = \frac{x_{4,t,z}}{A_z},$$

where A_z represents the land area of zipcode z in square miles, as recorded in the 2010 Census.[9] The time-dependent financial covariates are deflated to constant 2009 dollars.

The models specified below were fit in R using the GAMLSS package. The time periods covered differed for the five data states, with the Arizona data covering 2000-2015, the New Jersey data covering 2001-2015, and the California, Maryland, and Massachusetts data covering 2007-2015. Because of positive correlations between the data values for the same zipcodes in different years, the *p*-values generated by GAMLSS (shown in Appendix B) may be somewhat understated. Prior knowledge of the residential PV market, however, indicates that the selected variables are significant drivers of PV installations.

4.2. Hurdle Model Specification

The reduced hurdle model is a two-part model comprising logit and zero-truncated negative binomial components. Both components use covariates from the list above as independent variables. The logit model specification is as follows:

$$\lambda_{t,z} = \ln\left(\frac{\pi_{t,z}}{1 - \pi_{t,z}}\right) = \alpha_0 + \alpha_1 x_{1,t,z} + \alpha_2 x_{3,t,z} + \alpha_3 x_{4,t,z} + \alpha_4 x_{7,t} + \alpha_5 x_{8,t} + \alpha_6 y_{t-1,z}, \qquad (1)$$

where, for zipcode z during year t, $\pi_{t,z}$ represents the probability of observing at least one solar PV installation, and $y_{t-1,z}$ represents the number of new installations observed in the previous year. The model is fit using a binary dependent variable, which takes on values in $\{0,1\}$, with 0 indicating no observed installations and 1 indicating one or more

[7] Annual average national-level solar PV prices are from a report issued by the Lawrence Berkeley National Laboratory, available at http://emp.lbl.gov/publications/tracking-sun-vii-historical-summary-installed-price-photovoltaics-united-states-1998-20. The prices are adjusted to account for the 30% federal tax credit implemented in 2006.

[8] Annual average national-level mortgage interest rates are from the Federal Reserve Economic Data (FRED) system.

[9] The land area dataset is available at http://proximityone.com/cen2010_zcta_dp.htm.

installations. For the four zipcodes with extreme covariate values, the lagged dependent variable $y_{t-1,z}$ is omitted from the logit model.

The zero-truncated negative binomial component of the reduced hurdle model is specified as follows:

$$\zeta_{t,z} = \ln\left(y_{t,z} \middle| y_{r,z} > 0\right) = \beta_0 + \beta_1 x_{1,t,z} + \beta_2 x_{2,z} + \beta_3 x_{3,t,z} + \beta_4 x_{4,t,z} \\ + \beta_5 x_{7,t} + \beta_6 x_{8,t,z}\ . \tag{2}$$

Because of the conditioning on the dependent variable ($y_{r,z} > 0$), the zero-truncated model is fit using only the zipcode/year observations with one or more solar PV installations. Model diagnostics for both component models are given in Appendix B. To extract the fitted value (the expected number of installations) for zipcode z in year t, we compute

$$\hat{y}_{t,z} = \frac{e^{\hat{\lambda}_{t,z} + \hat{\zeta}_{t,z}}}{1 + e^{\hat{\lambda}_{t,z}}}, \tag{3}$$

where $\hat{\lambda}_{t,z}$ and $\hat{\zeta}_{t,z}$ are computed using the estimated coefficient vectors $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ from model equations (1) and (2) along with the covariates for zipcode z in year t.

4.3. Dynamic Model Coefficients

Coefficients for the hurdle model, as estimated from historical data, appear in Appendix B. As residential solar PV becomes more common and affordable, however, the effects of income, retail electricity rates, and social spillover (represented by the lagged dependent variable) are expected to decline. For the 2018 version of the NEMS, we therefore applied dampening factors to the coefficients for these variables. The factors ranged from 1 (at the beginning of the projection period) to 0.82 (40 years after the most recent year of historical data). The formula for the dampening factor for year t is

$$d_t = \frac{1}{(t - t_0)^{0.055}}, \tag{4}$$

where t_0 is the most recent year for which historical data are available. The constant 0.055 was chosen by expert judgement following a sensitivity analysis. The dampening factors appear in Figure 1.

4.4. Matching Target Zipcodes to Data Zipcodes

In order to apply the lagged hurdle model to the target zipcodes, we impute installation counts for each target zipcode from a data zipcode that is similar in the covariates x_1 through x_4 defined in subsection 4.1. We first average the time-dependent covariates for each zipcode for the years 2007 to 2015. For $i \in \{1,2,4\}$, let

$$\bar{x}_{i,z} = \frac{1}{9} \sum_{t=2007}^{2015} x_{i,t,z}.$$

Because the covariates are expressed in different units with substantially different magnitudes, we standardize them prior to computing Euclidian distances.

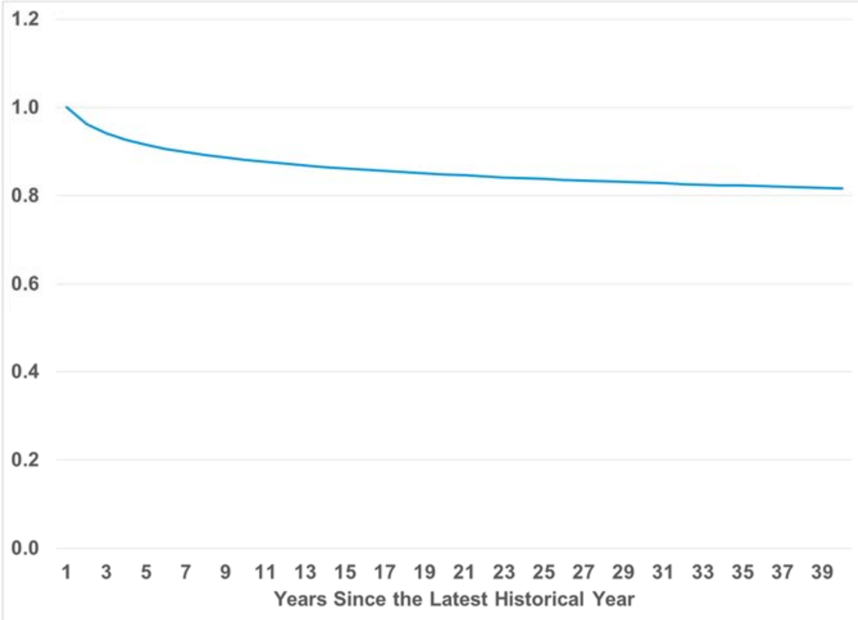

Figure 1: Dampening Factors for Coefficients of Income, Electricity Rate, and Spillover Effect

The standardization and matching process is done by state. Let s be a target state, and let d represent the collection of data states. We combine the zipcodes from s and d for the standardization. For $i \in \{1, \ldots, 4\}$ and $z \in s \cup d$, let $\mu_{i,s,d}$ and $\sigma_{i,s,d}$ be the mean and standard deviation, respectively, of $\bar{x}_{i,z}$, taken over the combined zipcodes $z \in s \cup d$. Let

$$\xi_{i,z} = \frac{\bar{x}_{i,z} - \mu_{i,s,d}}{\sigma_{i,s,d}}.$$

For each target zipcode $z_s \in s$ and each data zipcode $z_d \in d$, let

$$u(z_s, z_d) = \sum_{i=1}^{4} \left(\xi_{i,z_s} - \xi_{i,z_d}\right)^2.$$

We match the target zipcode z_s to the data zipcode z_d that minimizes $u(z_s, z_d)$. To impute a solar PV installation count $\tilde{y}_{t,z_s}$ for the target zipcode z_s in year t, we adjust the count from the matched zipcode z_d by the ratio of the fitted values from a reduced hurdle model that includes all independent variables shown in equations (1) and (2) except the lagged dependent variable. The imputed value is

$$\tilde{y}_{t,z_s} = y_{t,z_d} \left(\frac{\hat{y}_{t,z_s}}{\hat{y}_{t,z_d}}\right) \left(\frac{x_{4,t,z_s}}{x_{4,t,z_d}}\right),$$

where y_{t,z_d} is the actual installation count from zipcode z_d, and $\hat{y}_{t,z_s}$ and $\hat{y}_{t,z_d}$ are the fitted values from the reduced hurdle model. The second adjustment factor accounts for differences between the sizes (numbers of households) in the target and data zipcodes.

5. Results

Appendix B shows the estimated coefficients of the hurdle model specified in Section 4. These coefficients were used, along with the NEMS projections of the model covariates,[10] to project the numbers of solar PV installations. The national-level results for the Reference case presented in the *Annual Energy Outlook 2015* (*AEO 2015*) are shown in Figure 2. The *AEO* Reference Case is based on current policies and an assumed annual GDP growth rate of 2.4%. For details on all of the *AEO 2015* cases, see "Assumptions to *AEO 2015*," available at http://www.eia.gov/forecasts/aeo/assumptions/. Figure 2 shows that, as compared to the previous "payback model," the hurdle models project larger numbers of PV installations in the years following 2021.

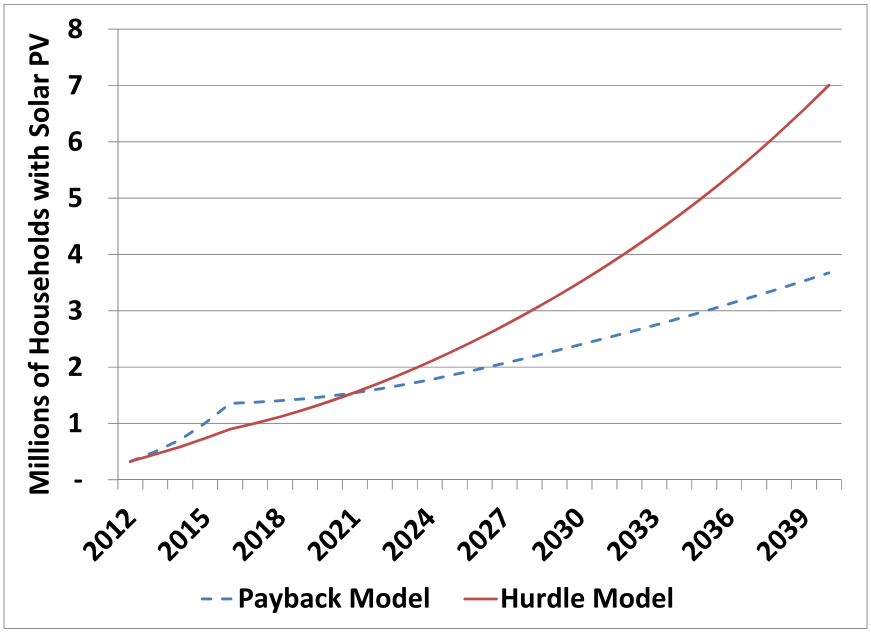

Figure 2: National Projections for the *AEO 2015* Reference Case, Hurdle vs. Payback Models

The 2016 leveling-off point in the payback model series in Figure 2 reflects the impact of the expiration of the federal investment tax credit for residential solar PV. The hurdle model series appears to be driven more by macroeconomic conditions and shows less impact of the tax credit expiration. Figure 3 shows the hurdle model results for the nine Census Divisions. Growth in all divisions is projected, with the Pacific and South Atlantic Divisions continuing to add the highest numbers of installations.

Figures 2 and 3 show series computed for the *AEO 2015*; these series include only roof-top solar installations on individual homes. Preliminary projections for the *AEO 2018* are substantially higher, because (a) they are adjusted to include community solar that is classified as residential and (b) they assume implementation of the Clean Power Plan, which provides incentives for renewable energy systems.

---

[10] Household income is not projected in the NEMS but was assumed to change in proportion to the projected changes in personal disposable income.

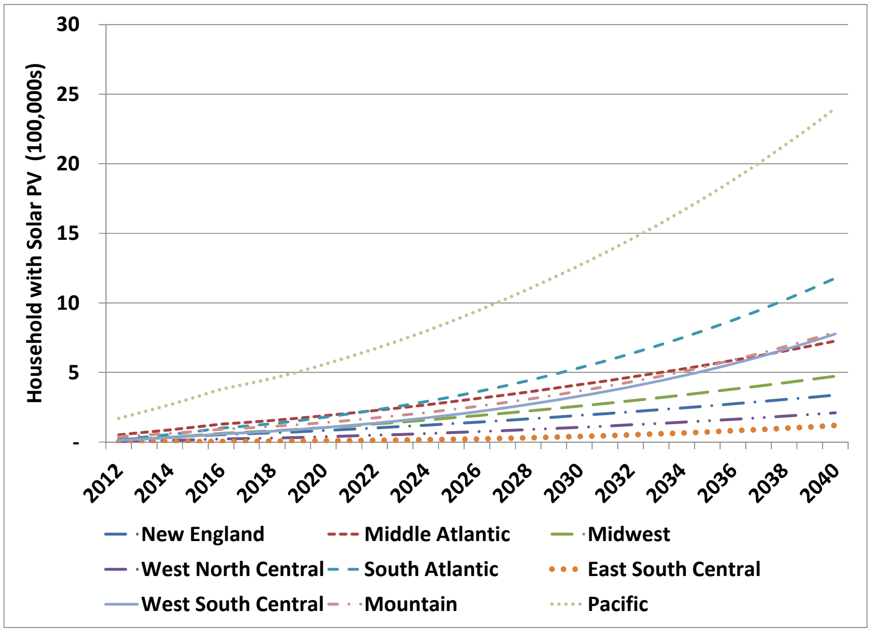

Figure 3: Regional Hurdle Model Projections for the *AEO 2015* Reference Case

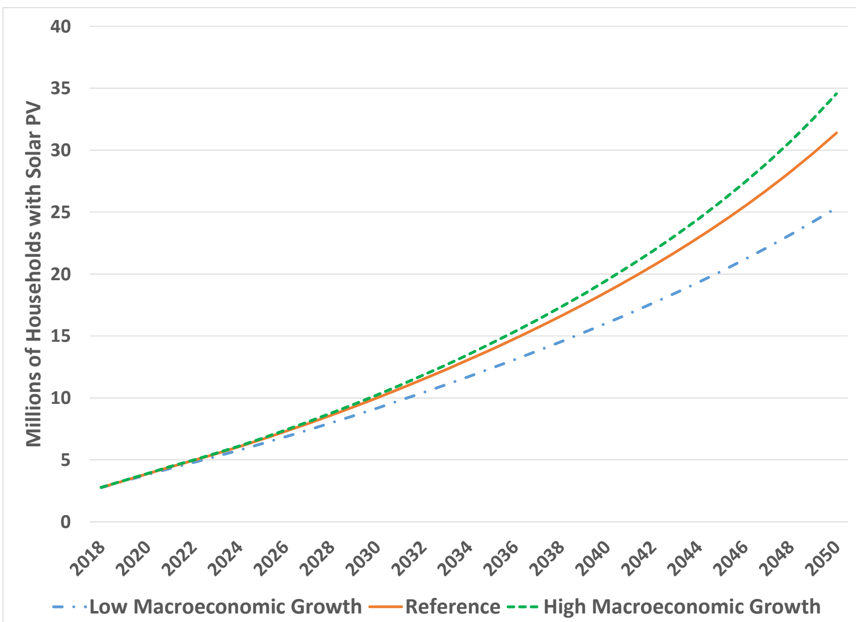

Figure 4: Preliminary High and Low Economic Growth Cases for the *AEO 2018*

Figure 4 shows preliminary projections, as of June 2017, for the *AEO 2018* Reference Case, along with those of the High Economic Growth and Low Economic Growth Cases. These cases assume higher and lower annual growth in the gross domestic product (GDP), respectively, as compared to the Reference Case. Household income, interest rates, and numbers of households are all affected by the differing macroeconomic assumptions, and the inclusion of these variables in the hurdle model causes the model projections to reflect substantially larger numbers of PV installations for the High versus the Low Economic Growth Case.

Figure 5 shows the preliminary hurdle model projections for the *AEO 2018* Reference Case along with those for the High and Low Oil and Gas Price Cases. The High and Low Oil and Gas Price Cases assume higher and lower values of the Brent crude oil price by 2050, as compared to the Reference Case. These cases reflect the macroeconomic effects of high and low oil and natural gas prices. As compared to the low oil price case, the high oil price case has the following projected differences:

a) higher electricity prices;
b) higher inflation, as measured by the GDP price deflator;
c) higher interest rates (10 year treasury note); and
d) lower real personal disposable income.

The higher electricity prices in the High Oil Price case would increase the incentive for homeowners to install solar PV systems. Higher inflation, however, would increase the prices of residential PV systems, and higher interest rates would increase the cost of financing the investment. The decrease in disposable income would also tend to decrease the number of PV installations. Figure 5 shows that, according to the hurdle models, the combined effect of (a) through (d) above is essentially negligible, as the effects of (b) through (d) counteract the effect of (a).

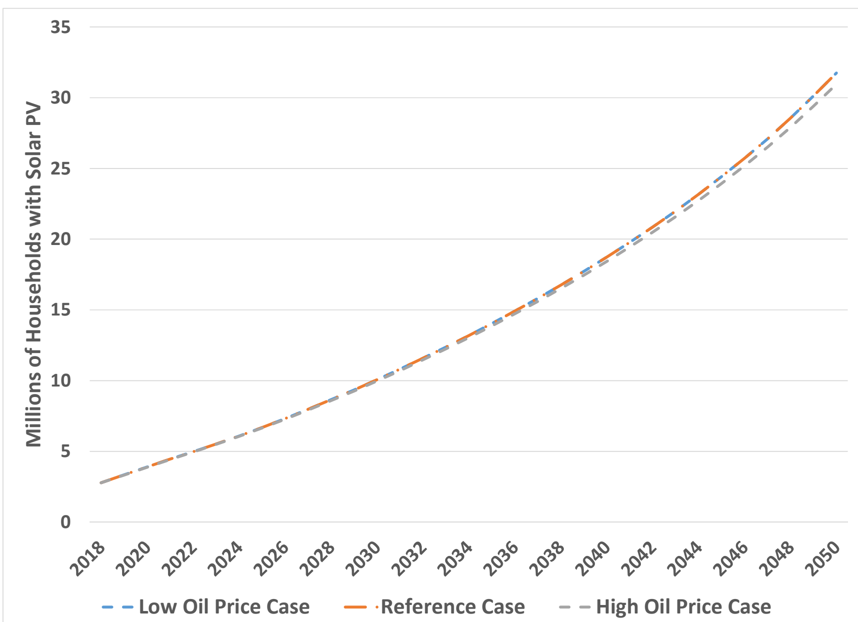

Figure 5: Preliminary High and Low Oil and Gas Price Cases for the *AEO 2018*

6. Future Research

EIA expects to further develop the hurdle models described above by incorporating additional data on PV installations. Data from additional states will be incorporated into the models as they become available, and the hurdle model projections will continue to be benchmarked to historical estimates.

The hurdle models cannot separate out the effects of different state policies from the effects of different levels of income, solar irradiation, etc., in the data states. Although the NEMS applies adjustments to account for some state-level policy differences, these capabilities may be enhanced in the future by incorporating the results of studies focused on state-level policy effects. NREL, for example, has compiled the results of several studies on its website (http://www.nrel.gov/analysis/policy_state_local.html).

Acknowledgements

Thanks are due to Steve Wade and Kevin Jarzomski of EIA for their help in preparing the projections graphed in the figures in this paper.

References

[1] Memorandum from Janice Lent to Steven Wade on "Research on the Use of American Community Survey Data to Project Residential Solar PV Penetration," February 12, 2015 (internal memorandum available upon request).

[2] Bilder, Christopher R. and Loughin, Thomas M. (2015). *Analysis of Categorical Data with R.* CRC Press, New York, pp. 61-140.

[3] Hernandez, Mari (2013). "Solar Power to the People: The Rise of Rooftop Solar Among the Middle Class." Center for American Progress, October 21, 2013.

[4] Hilbe, Joseph M. (2011). *Negative Binomial Regression,* 2nd edition. Cambridge University Press, UK.

[5] MIT Energy Initiative (2015). "The Future of Solar Energy: An Interdisciplinary MIT Study led by the MIT Energy Initiative," available at https://mitei.mit.edu/futureofsolar.

[6] Rothfield, Emily (2010). "Solar Photovoltaic Installation in California: Understanding the Likelihood of Adoption Given Incentives, Electricity Pricing and Consumer Characteristics." Honors Thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University. Durham, North Carolina.

[7] Seba, Tony (2014). *Clean Disruption of Energy and Transportation: How Silicon Valley Will Make Oil, Nuclear, Natural Gas, Coal, Electric Utilities and Conventional Cars Obsolete by 2030.* Published by Tony Seba; Beta edition; May 20, 2014.

[8] Wilcox, Stephen (2012). "National Solar Radiation Database 1991–2010 Update: User's Manual." Technical Report, NREL/TP-5500-54824, August, 2012, available at http://www.nrel.gov/docs/fy12osti/54824.pdf.

[9] Wirth, Harry (2015). "Recent Facts about Photovoltaics in Germany," published by Fraunhofer ISE and available at https://www.ise.fraunhofer.de/en/publications/veroeffentlichungen-pdf-dateien-en/studien-und-konzeptpapiere/recent-facts-about-photovoltaics-in-germany.pdf

Appendix A: Estimating Zipcode-level Electricity Rates

The zipcode-level electricity rate estimates are computed by combining average state-level rates from form EIA-861 ("Annual Electric Power Industry Report") with zipcode-level estimates for February, 2011. The zipcode level estimates were developed by NREL using inputs from EIA and Ventyx Research, Inc. The estimates and accompanying documentation are available at the following link:
http://catalog.data.gov/dataset/u-s-electric-utility-companies-and-rates-look-up-by-zipcode-feb-2011-57a7c.

The goal of the estimation is to use the NREL estimates to capture some of the variation in electricity rates by zipcode within each state, while calibrating the zipcode-level estimates to the state-level averages from form EIA-861. For each state s and year t, let $p_{t,s}$ denote the state-level average electricity rate from form EIA-861. For each zipcode $z \in s$, let p_z denote the zipcode-level rate from the NREL data, computed as a simple average of the rates for investor-owned and non-investor-owned utilities. (The rates for the two types of utilities differed little for most zipcodes.) Let $x_{t,z} = x_{4,t,z}$, the estimated number of households in zipcode z in year t, from the ACS and decennial census data. Let

$$\bar{p}_{t,s} = \frac{\sum_{z \in s} p_z \, x_{t,z}}{\sum_{z \in s} x_{t,z}}.$$

We estimate the electricity rate for zipcode z in year t as

$$\hat{p}_{t,z} = p_z \left(\frac{p_{t,s}}{\bar{p}_{t,s}} \right).$$

Averaging the zipcode-level estimated rates within the state, weighted by the estimated numbers of households, returns the state-level average rate ($p_{t,s}$) from form EIA-861.

Appendix B: Model Diagnostics

<u>Logit Component</u>

```
Family:  c("BI", "Binomial")

Call:
gamlss(formula = AnyInstalls ~ Households + PopDensity +
    Income + ElecRate + MonthlyPayment + Lag1_Installs,
    family = BI(mu.link = "logit"), data = CANJAZPlus,
    method = CG())

Fitting method: CG()

Mu link function:  logit
Mu Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.965e+00  7.794e-04 -3804.1   <2e-16 ***
Households      7.742e-05  2.650e-08  2921.8   <2e-16 ***
PopDensity     -1.076e-04  4.253e-06   -25.3   <2e-16 ***
Income          1.051e-05  2.760e-09  3807.4   <2e-16 ***
ElecRate        1.371e-01  3.547e-05  3864.5   <2e-16 ***
MonthlyPayment -2.390e-02  6.018e-06 -3972.0   <2e-16 ***
Lag1_Installs   2.891e-01  7.247e-05  3989.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-----------------------------------------------------------------
No. of observations in the fit:  30468
Degrees of Freedom for the fit:  7
      Residual Deg. of Freedom:  30461
```

```
                      at cycle:  8

Global Deviance:     28855.56
            AIC:     28869.56
            SBC:     28927.83
```

Zero-truncated Negative Binomial Component

```
Family:  c("NBIlefttr", "left truncated Negative Binomial type
I")

Call:  gamlss(formula = Installs ~ Households + PopDensity +
    Income + Insol + ElecRate + MonthlyPayment, family =
NBIlefttr,
    data = CANJAZPlusPos)

Fitting method: RS()

Mu link function:  log
Mu Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.689e+00  1.296e-01  -13.04   <2e-16 ***
Households      1.290e-04  2.111e-06   61.12   <2e-16 ***
PopDensity     -1.193e-04  5.074e-06  -23.52   <2e-16 ***
Income          9.722e-06  3.630e-07   26.78   <2e-16 ***
Insol           4.652e-01  1.473e-02   31.59   <2e-16 ***
ElecRate        1.101e-01  4.312e-03   25.53   <2e-16 ***
MonthlyPayment -8.428e-02  1.171e-03  -72.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-----------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4852     0.0205   23.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-----------------------------------------------------------------
No. of observations in the fit:  17614
Degrees of Freedom for the fit:  8
      Residual Deg. of Freedom:  17606
                      at cycle:  9

Global Deviance:     118281.9
            AIC:     118297.9
            SBC:     118360.1
```