Using R for Advanced Statistical Modeling in the National Energy Modeling System

Janice Lent, Chief Statistician

U.S. Energy Information Administration

October 24, 2018 | Workshop on Government Advances in Statistical Programming, Washington, DC

Overview

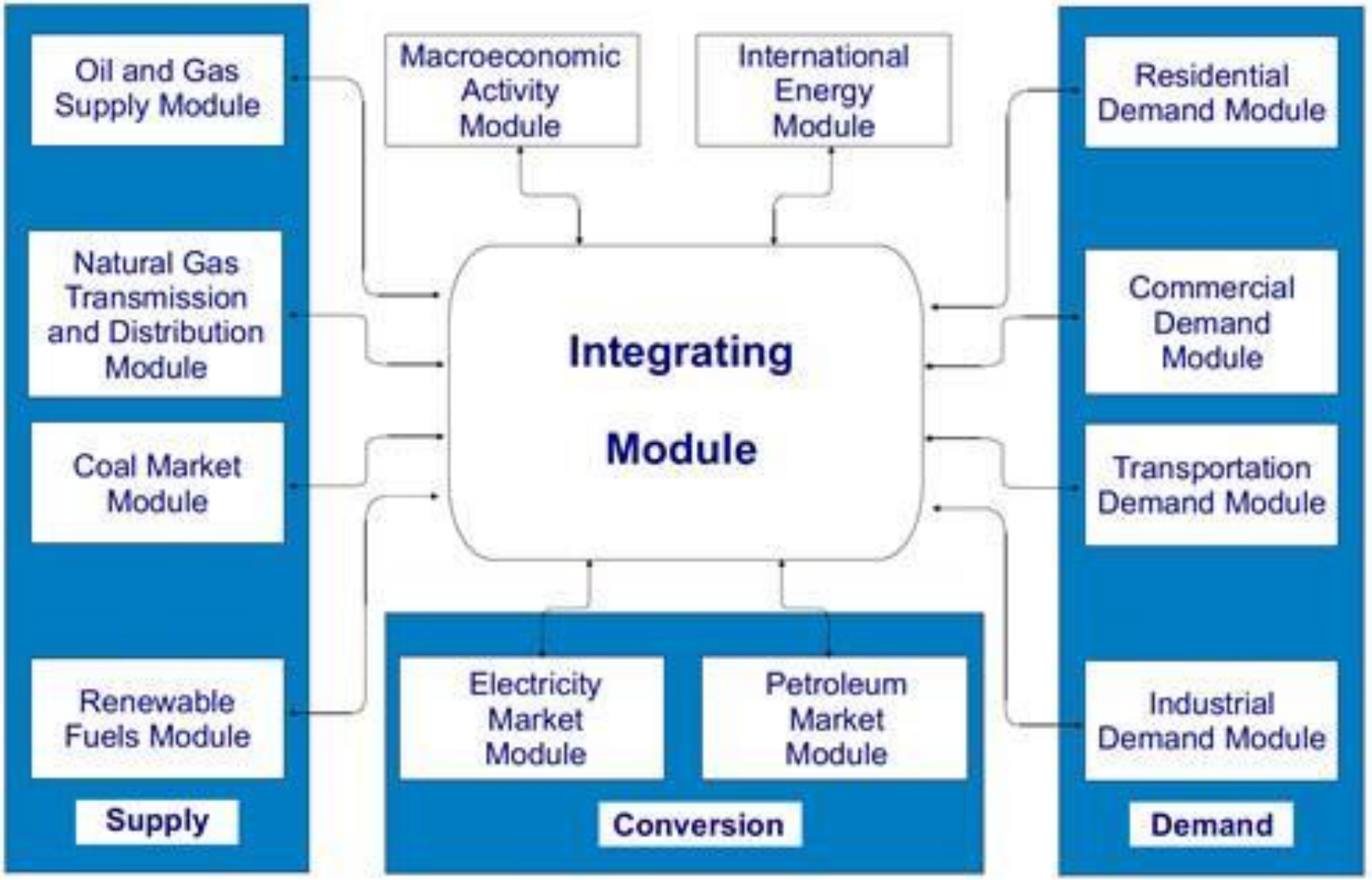
- Background: EIA's National Energy Modeling System (NEMS)
- Using R in the NEMS
- Applications
  - Hurdle models to project residential solar photovoltaic system installations
  - Bayesian dynamic linear models for spot prices of hydrocarbon gas liquids
  - Calibrating long-term projections to volatile short-term projections

EIA's National Energy Modeling System (NEMS)

- EIA's long-term projection system, the NEMS projects annual U.S. energy production, consumption, prices, etc. for EIA's *Annual Energy Outlook* (AEO).
- Inputs to the NEMS
  - A vast collection of EIA and non-EIA datasets, including global economic and demographic information
  - Classic economic assumptions regarding the effects of long-term supply and demand pressures
  - Specific assumptions regarding macroeconomic conditions and technological advances, as specified in NEMS "cases."

Examples of NEMS "cases"

- Reference case—status quo, assuming moderate GDP growth, relative to historical averages
- High and low macroeconomic growth cases
- High and low oil and natural gas price cases
- High and low technological advancement cases
- Special policy-based cases, e.g., cases representing the effects of the Clean Power Plan proposed in 2014.


Oil and Gas Supply Module
Natural Gas Transmission and Distribution Module
Coal Market Module
Renewable Fuels Module
Supply
Macroeconomic Activity Module
International Energy Module
Integrating Module
Electricity Market Module
Petroleum Market Module
Conversion
Residential Demand Module
Commercial Demand Module
Transportation Demand Module
Industrial Demand Module
Demand

Using R in the NEMS

- The NEMS base code is written in Fortran.
- The NEMS invokes R by sending a call to the command prompt.
- R opens and runs the script indicated in the Fortran command.
- R reads and writes comma-delimited files to exchange data with the NEMS base code.

Projecting the growth of residential solar photovoltaic (PV) penetration in the NEMS

Previous NEMS method of projecting growth of residential solar PV

- Projection strata were defined by cross-classifying housing units based on variables from EIA's Residential Energy Consumption Survey (RECS) and data on solar irradiation from the National Renewable Energy Laboratory.
- For each projection stratum, a cash flow analysis produced an estimated "payoff time" for a residential solar PV installation. Projected installations were based on payoff time.
- Projections were driven mainly by projected PV prices and failed to reflect social momentum and the effects of macroeconomic assumptions in different NEMS cases.

An econometric modeling approach

- Starting with the 2017 version of the NEMS, EIA implemented a new econometric approach for modeling residential solar PV.
- Reflects social influences and macroeconomic assumptions associated with specific NEMS cases.
- Incorporates data from new sources.
- Uses zipcodes as projection strata.

Use of "big data" sources

- Project-level data on PV installations
  - Some state governments publish project-level databases
  - National Renewable Energy Laboratory's (NREL's) Open PV Project database (voluntarily-reported data on PV installations)
- Completeness of the data is uncertain.
- In 2014, EIA began publishing estimates of historical residential solar PV electricity generating capacity, providing historical data to which projections may be calibrated.

Residential solar PV installation data

- We aggregate the project-level data to the zipcode level for each installation year.
- The observations are zipcode/year combinations.
- We calibrate the model-based estimates to recent historical data at the state level.

Model covariates

- Because we wish to use the model for long-term projections, we limit our model covariates to items that are projected in the NEMS.

- Model covariates
  - Median income and number of households, by zipcode, from the American Community Service (ACS)
  - Population density (households per square mile, used as a proxy for roof area) from the ACS and Postal Service
  - Zipcode-level residential electricity rates, estimated from EIA and NREL data

Model covariates (2)

- Model covariates (continued)
  - Annual average solar irradiation (kilowatt hours per square meter per day) from NREL
  - Monthly payment per watt of installed PV capacity, estimated using PV price data from the Lawrence Berkeley National Laboratory and annual average mortgage interest rates from the Federal Reserve Economic Data (FRED) system
  - A one-year lagged dependent variable, to represent local social momentum.
- The NEMS projects the covariates at the census division level. We assume change is constant within divisions.

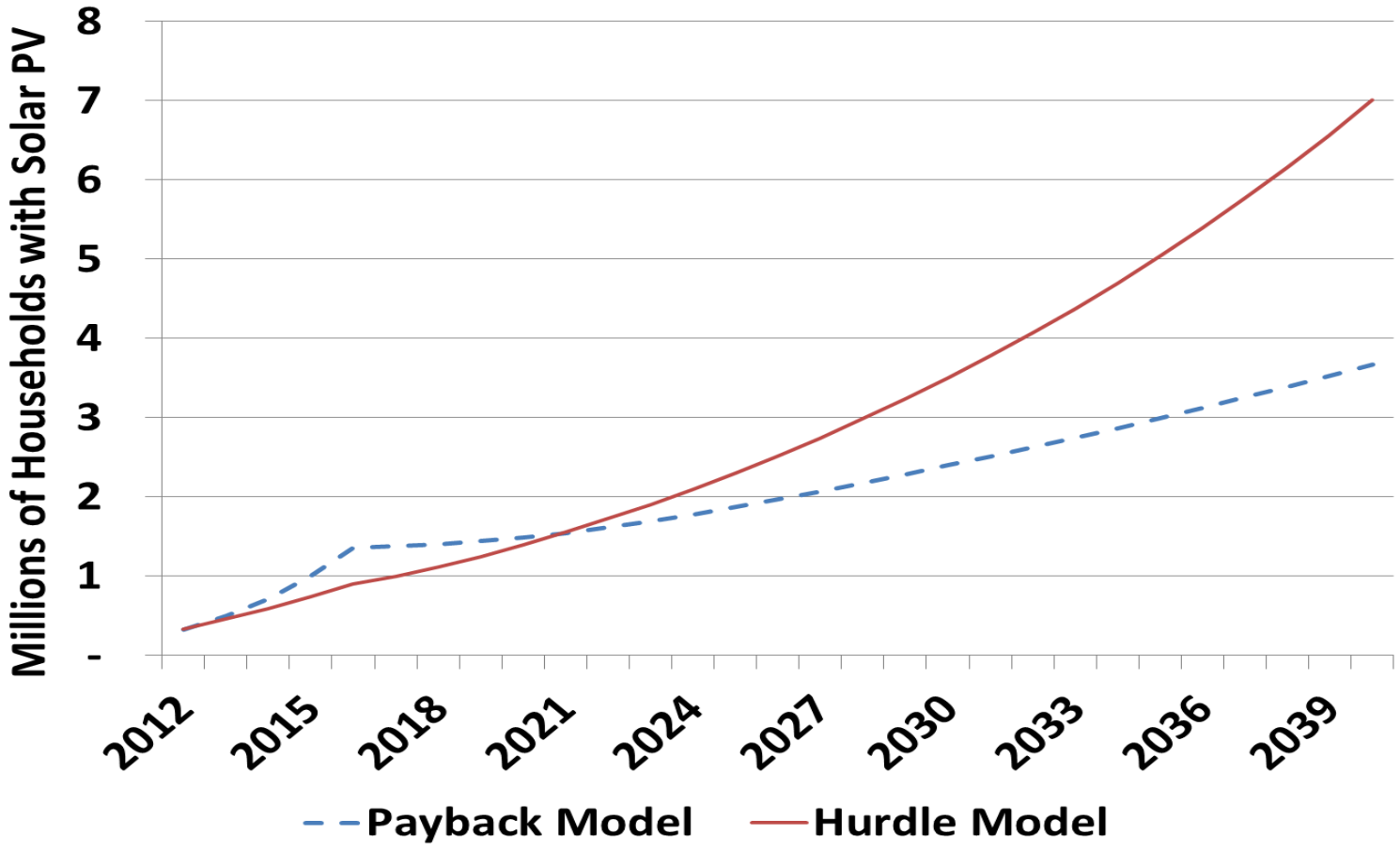
Model fitting algorithm

- The zero-hurdle model comprises a logistic regression model combined with a negative binomial count model.

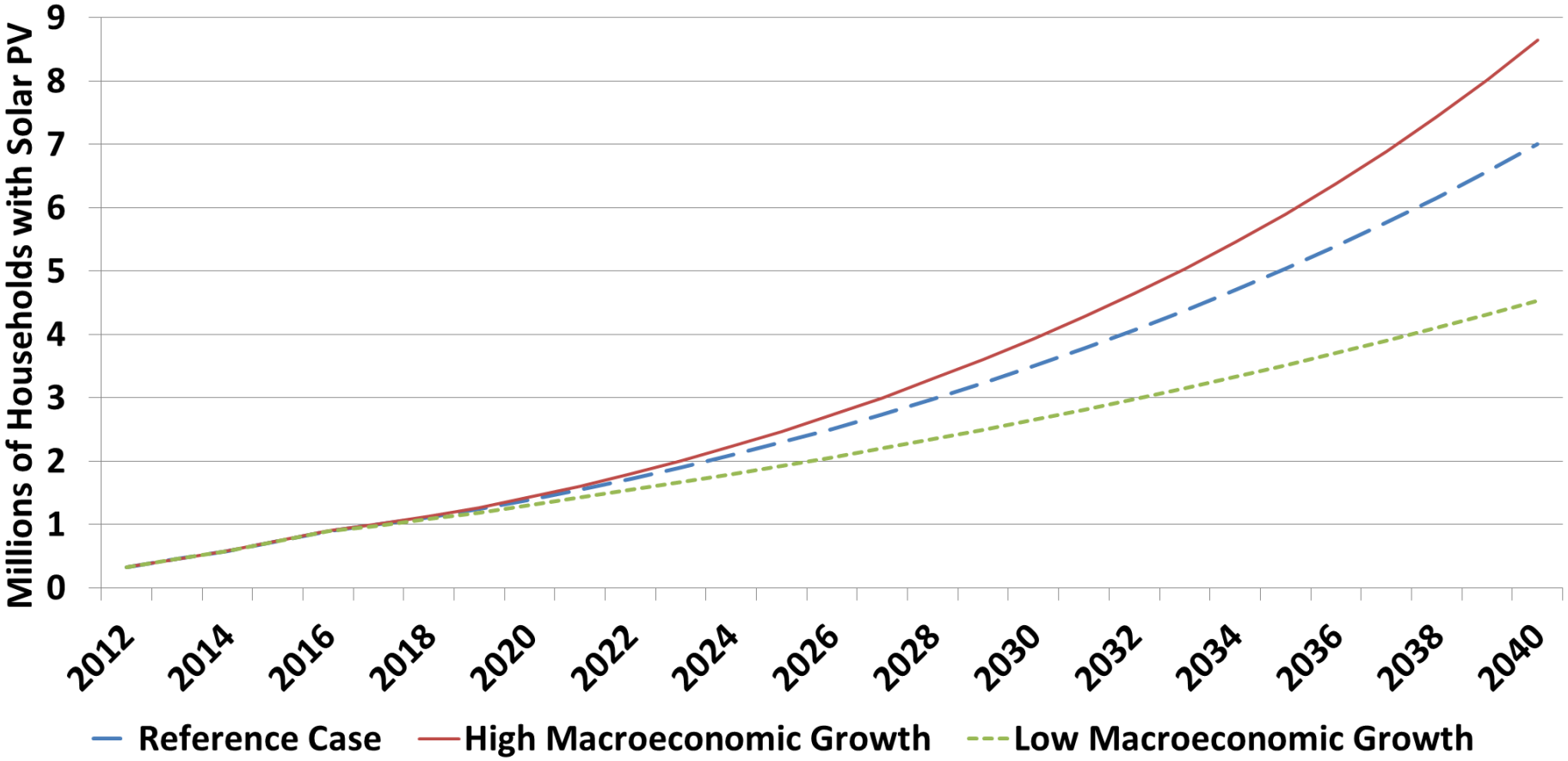
- We fit the component models using functions in the contributed R package **gamlss** (Generalized Additive Models with Location, Scale, and Shape).

- By default, the **gamlss** package uses the Rigby-Stasinopolis (RS) method, a weighted iterative maximum likelihood algorithm, to fit the non-linear models.

Hurdle model vs. payback model – reference case



Hurdle model for macroeconomic cases



Advantages of the new model

- Uses additional data sources, including ACS data and project-level datasets of PV installations.
- Reflects social momentum at the zipcode level.
- Reflects macroeconomic and other assumptions associated with different NEMS cases.

Projecting prices of hydrocarbon gas liquids (HGLs)

- Hydrocarbon gas liquids include a range of substances produced during natural gas processing and petroleum refining.
- Advances in horizontal drilling and hydraulic fracturing ("fracking") technologies have recently led to record high levels of HGL production in the United States.
- The "shale gas boom" has affected the market dynamics of HGLs.
- The most plentiful HGLs are propane and ethane.

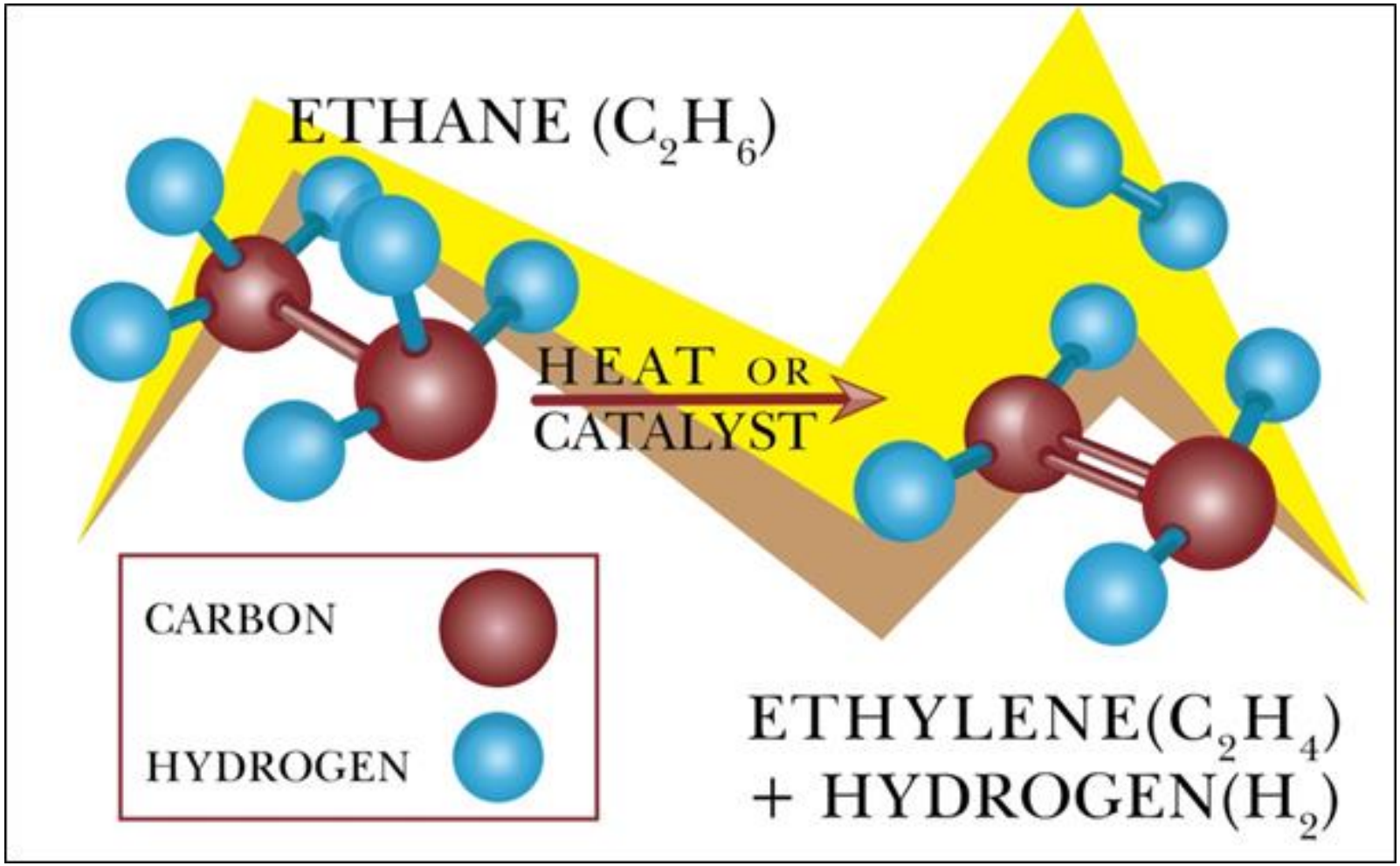
Market dynamics for propane

- Propane (C_3H_8)
  - Vaporizes at -44° F (-42° C), allowing for easy shipment in liquid form.
  - Variety of uses, including heating, in industrial, commercial, transportation, and residential sectors.
  - Domestic demand and exports are expected to expand with increasing supply.

Market dynamics for ethane

- Ethane (C_2H_6)
  - Vaporizes at -126° F (-88° C) and is usually transported by pipeline, limiting export potential (compared to propane), though exports by ship begun in recent years.
  - Increasing supplies from the Marcellus/Utica shale basin have decreased prices.
  - Used only by the petrochemical industry for “cracking” to produce ethylene.

The ethane cracking process



Propane and ethane prices

- Prior to horizontal drilling, most domestically produced propane came from oil refineries, as a byproduct of gasoline and diesel fuel.
- The previous NEMS model for propane prices was an ordinary least squares regression model with crude oil prices as the sole explanatory variable.
- EIA previously published no projected prices for ethane.

Changing price dynamics of HGLs

- Most domestically produced propane and ethane now comes from natural gas wells.
- Propane and ethane prices tend to track both oil and natural gas prices.
- When oil prices are rising faster (or decreasing more slowly) than natural gas prices, the HGL prices tend to track natural gas prices more closely, and conversely.
- We needed a model that would reflect the changing relative influences of Brent oil prices and Henry Hub (HH) natural gas prices on HGL prices.

Goals of hydrogen gas liquids spot price projection for the NEMS

- Use dynamic models to describe evolving relationships between energy spot prices.
- Systematize the use of expert judgment in the long-term projections, making it more transparent and more easily automated.
- Extend dynamic linear modeling techniques, developed for short-term forecasting, to the case of long-term forecasting.

Ordinary vs. dynamic linear models

- In a standard linear model, the parameters are fixed.

$$y_t = \theta_1 + \theta_2 x_t + \varepsilon_t, \qquad \varepsilon_t \sim \Phi(0, \sigma^2).$$

- A dynamic linear model (DLM) is a time-dependent parameter model with an observation equation:

$$Y_t = F_t \theta_t + v_t, \qquad v_t \sim \Phi(0, V_t),$$

and a state or system equation which describes the evolution of the parameters over time:

$$\theta_t = G_t \theta_{t-1} + w_t, \qquad w_t \sim \Phi(0, W_t).$$

Bayesian Dynamic Linear Models (DLMs)

- DLM computations are performed recursively, incorporating new data via a Bayesian updating process.
- For long-term projections, DLMs provide the flexibility needed to represent expected changes in energy market dynamics.
- We use a "multi-process" DLM, which employs a combination of models to represent the complexities of the HGL market price dynamics.

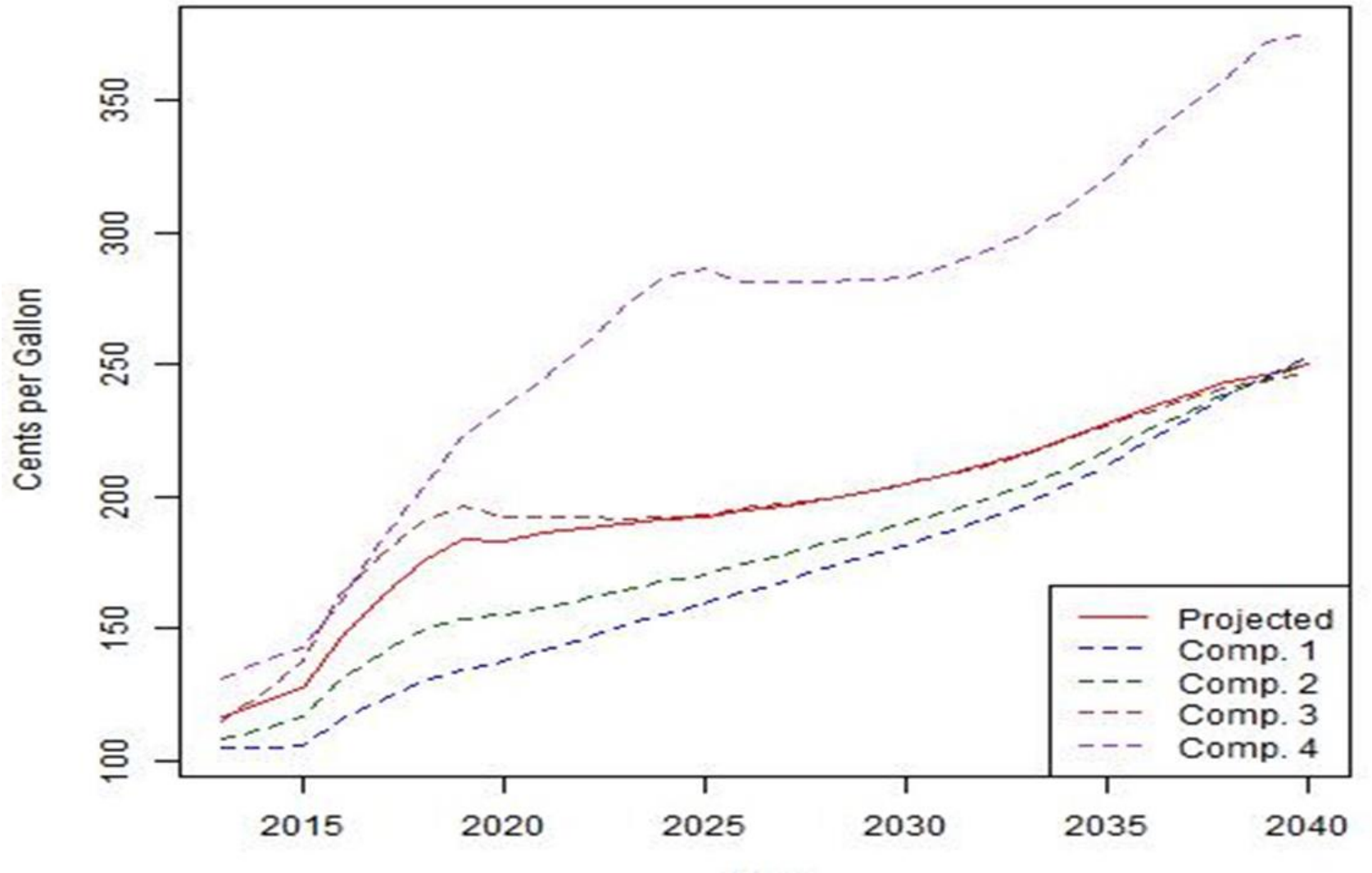
NEMS joint propane/ethane price model

- Bayesian multi-process model, incorporating four component models:
  - Three Bayesian DLMs with Brent oil prices and HH parameters that change at different rates (slow, moderate, and fast) over the projection period
  - A random walk DLM with parameters based on historical data for the Brent, HH, and the following supply and demand variables:
    - Ethane production
    - Propane production
    - Organic chemical demand

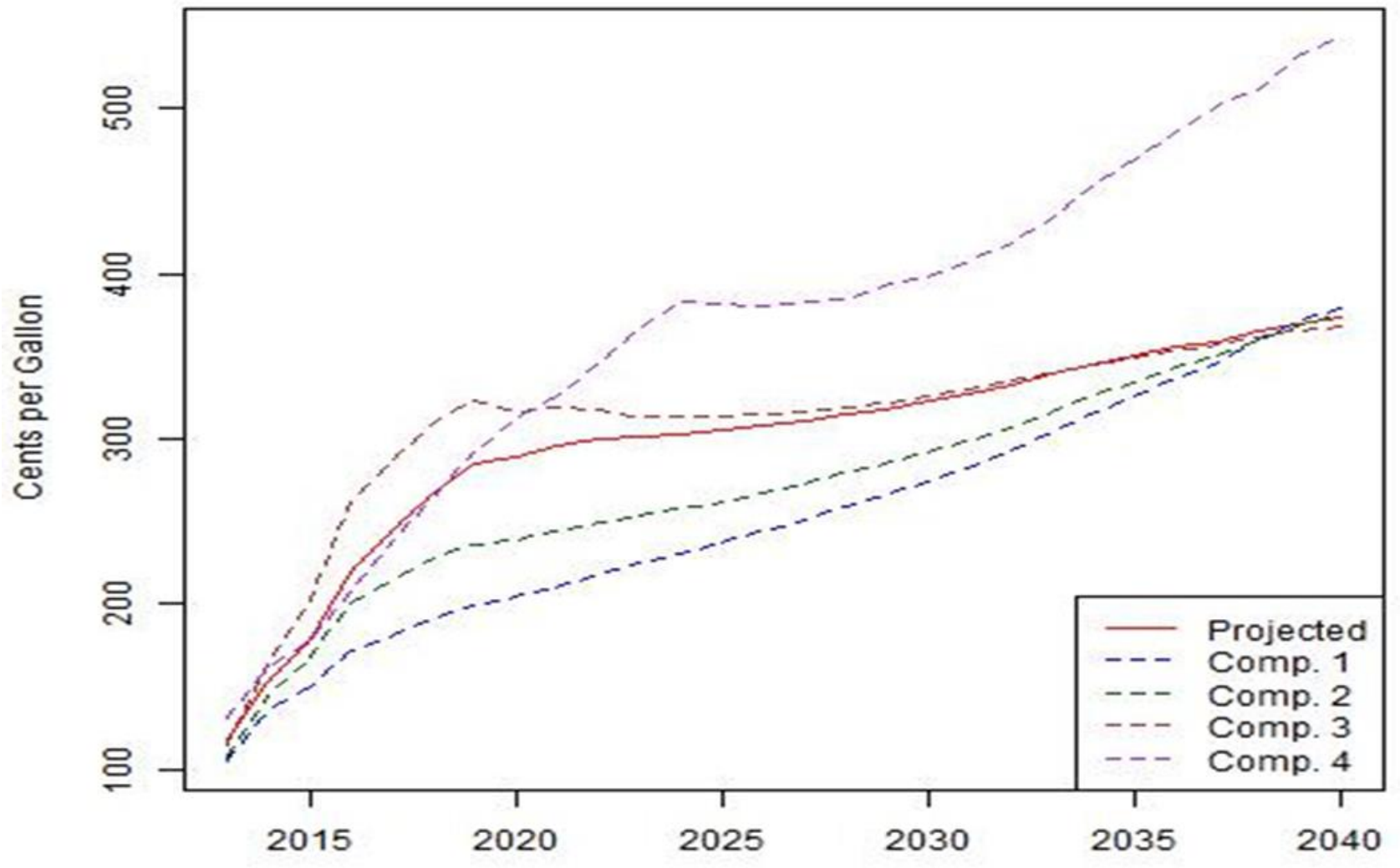
NEMS joint propane/ethane price model (2)

- The component weights change over the projection period, gradually reducing the influence of the historical data and increasing the influence of the Bayesian components.
- Hyperparameters control the evolution of the component weights over the projection period and can be customized for different NEMS cases (sets of assumptions) considered in EIA's *Annual Energy Outlook*.
- The joint propane/ethane price model is programmed using the **dlm** package in R. EIA contracted with the package author, Giovanni Petris, to assist with the programming.

Propane price projections: Reference case



Propane price projections: High oil price



Advantages of the new model

- Through Bayesian priors, the model formalizes the incorporation of expert judgment into the long-term projections.
- The dynamic model reflects the changing dynamics of the HGL markets.

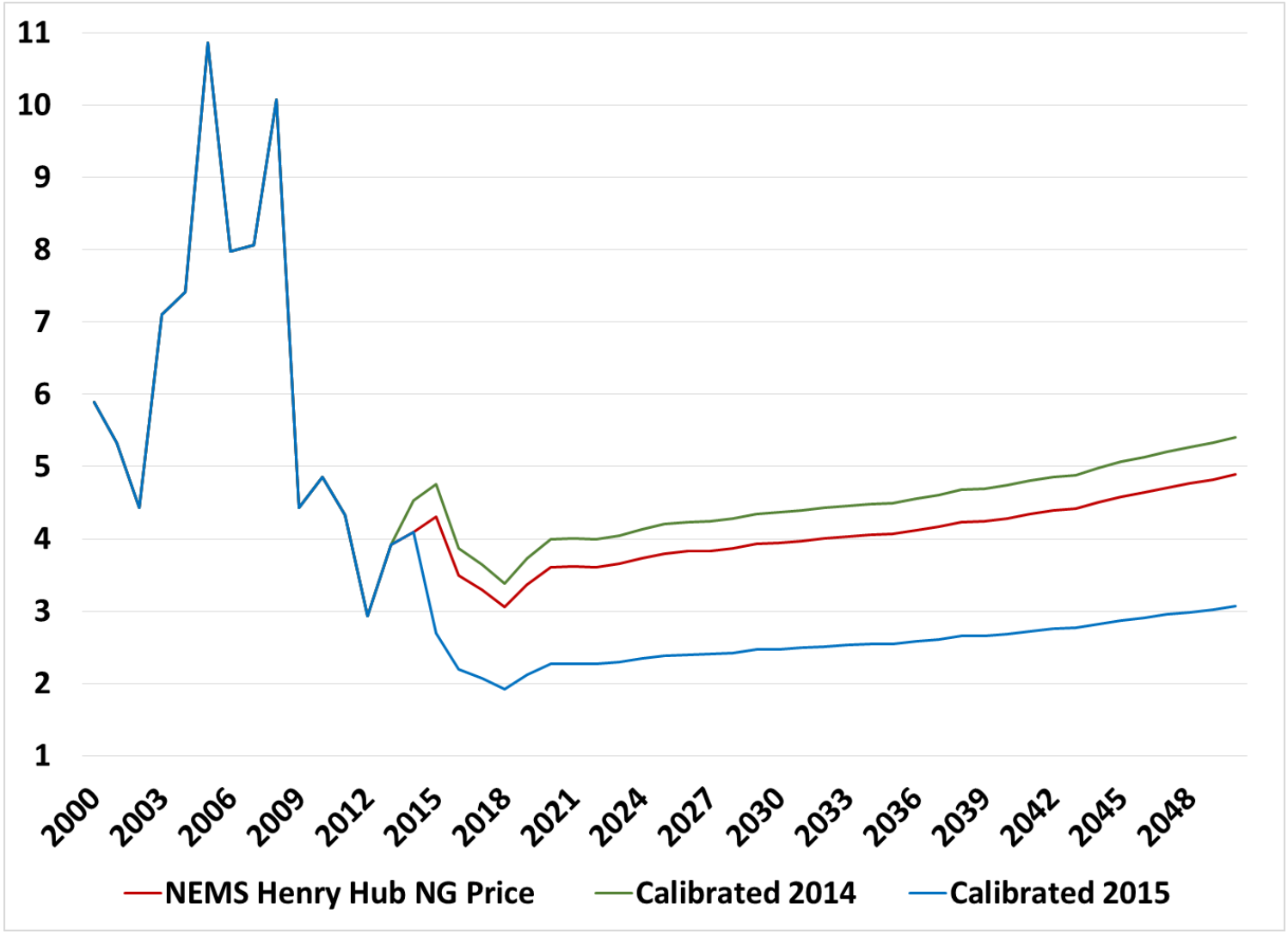
Using R for time-series calibration

- In addition to the long-term NEMS projections, EIA publishes monthly short-term projections.
- The monthly short-term projection period is one to two years. At the end of each year, the projection period is extended another year.
- The modeling system is called the *Short-Term Energy Outlook* (STEO) system.
- The annual long-term projections from the NEMS are calibrated to agree with the annualized STEO projections for the early years of the long-term projection period.

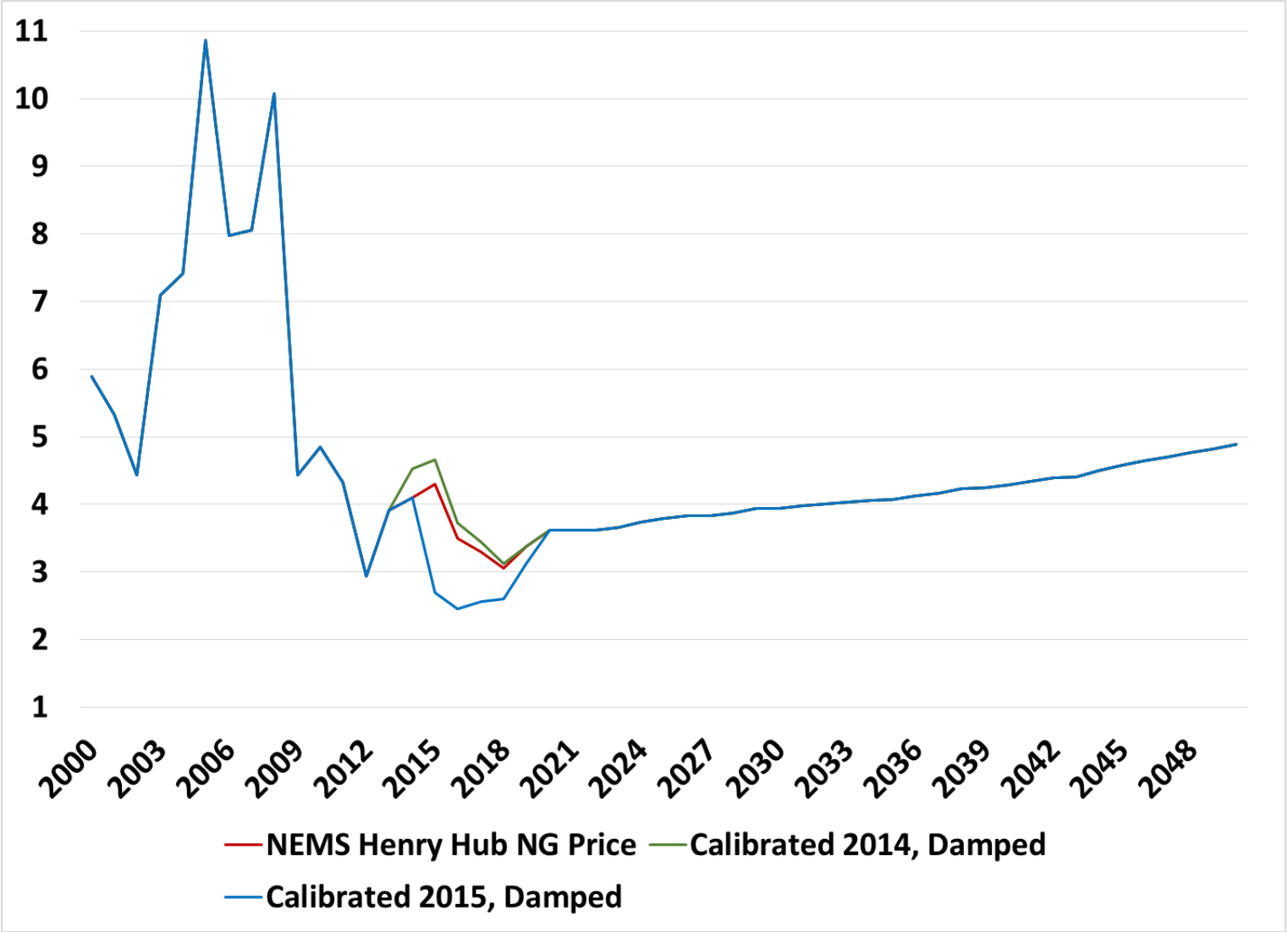
Calibration to short-term projections

- Because the historical data series for many energy-related statistics are volatile, the short-term projections often change substantially from year to year.
- The adjustment factors used to calibrate the long-term series are therefore sensitive to the choice of calibration year.

Example: Henry Hub Natural Gas Prices (dollars per MMBtu)



Example: Henry Hub Natural Gas Prices – with damping



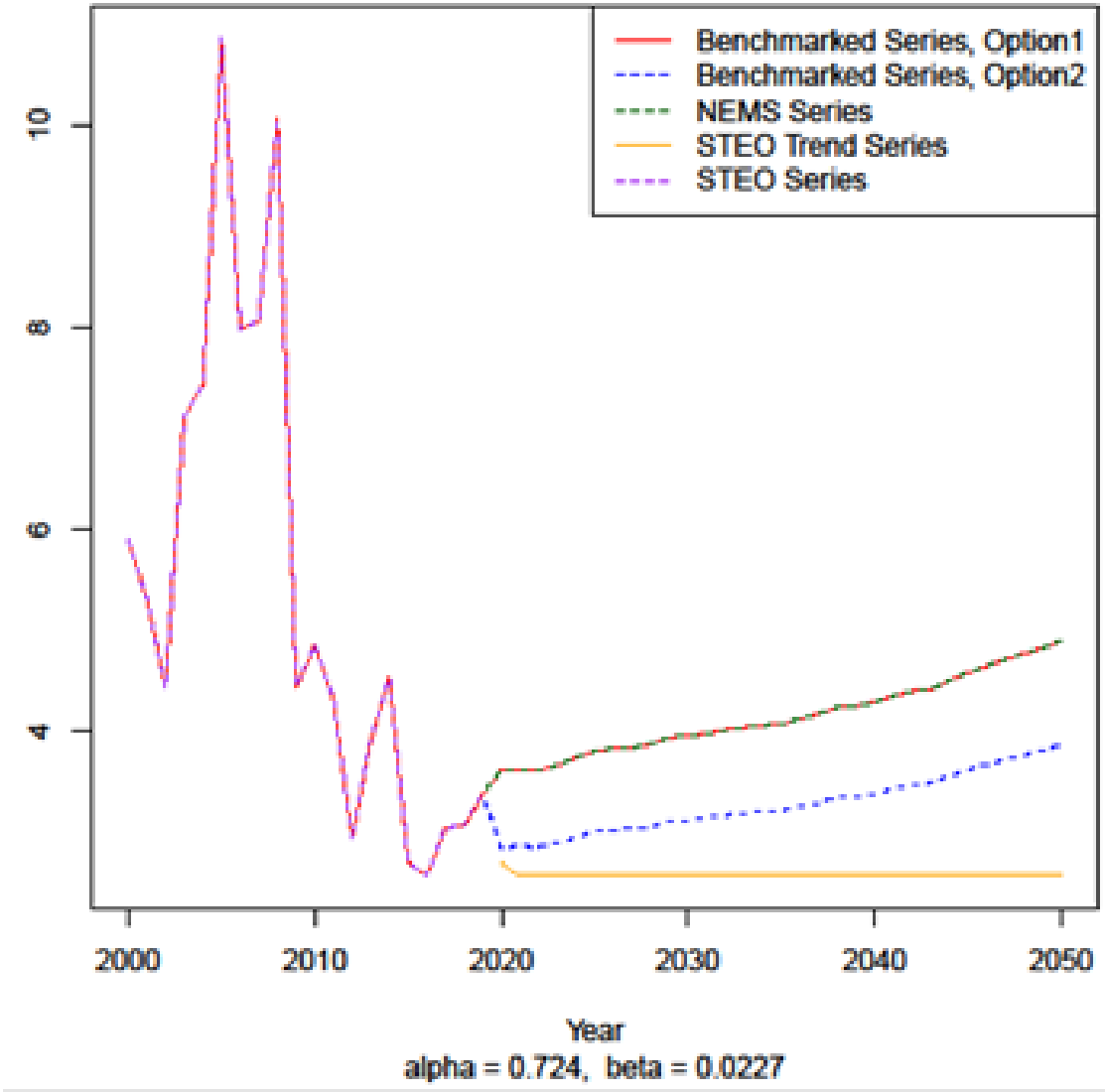
New calibration system

- The new calibration methods offer the option of calibrating to the *trend* of the short-term series, rather than to a particular value in the series.
- Rather than simply applying ratio adjustments, the new calibration methods involve linearly splicing time series, i.e., smoothing breaks in the series by inserting lines.
- The methods are implemented in R using the **zoo** package and the Holt-Winters (HW) function in base R.

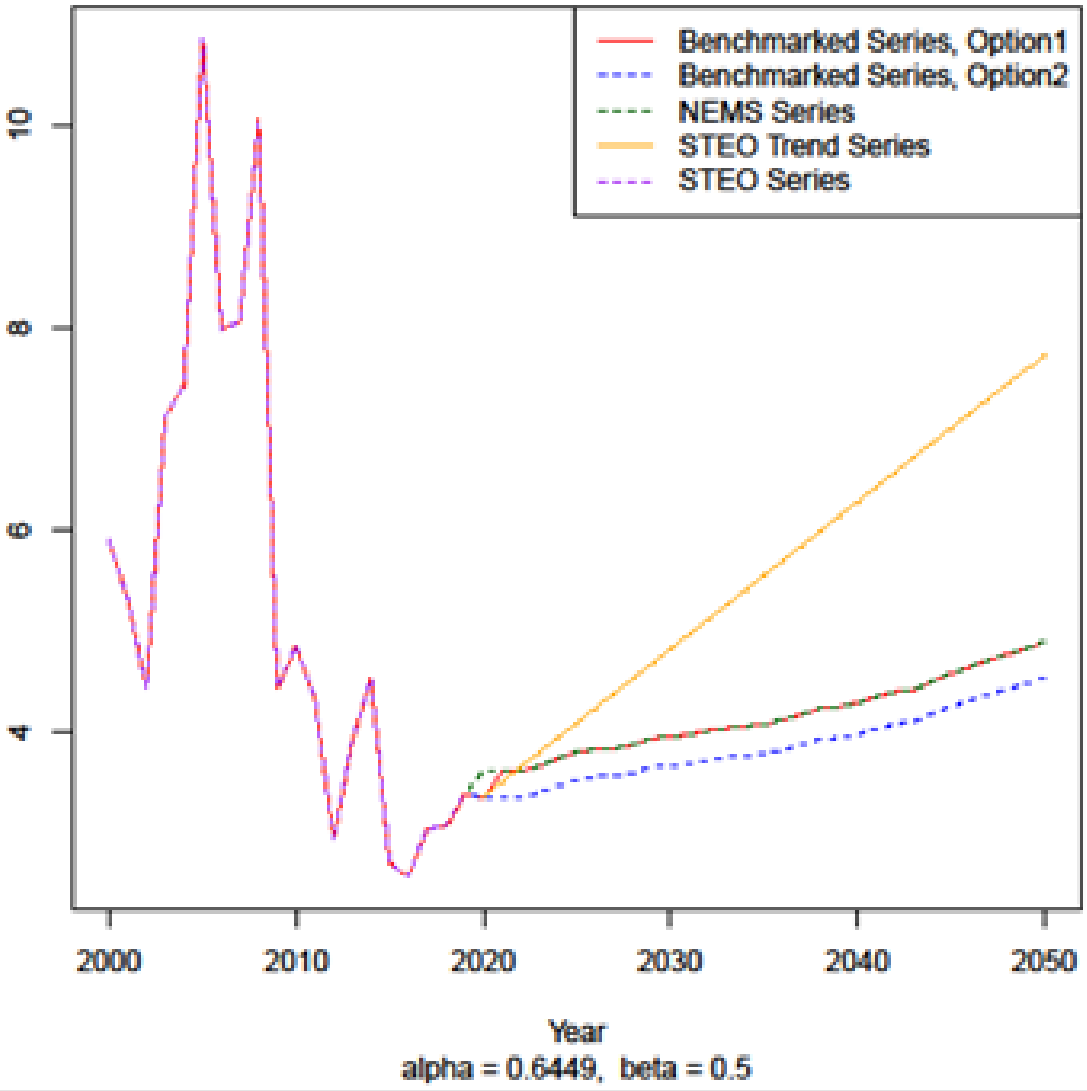
User inputs to the calibration methods

- Users may customize the methods by setting values for several parameters.
- Example:
  - The HW smoothing involves two parameters, α and β, which take values in the half-open interval $(0, 1]$.
  - Users may specify values for α and β. Lower values cause earlier data points in the series to be weighted more heavily, while parameter values near 1 give the more recent data points more weight.

Henry Hub NG price with a default β



Henry Hub NG price with a customized β



Benefits of the new calibration system

- By calibrating to the overall trend of the short-term series, the methods reduce the sensitivity of the long-term calibrated series to the level of the final year of the short-term projection.
- Through user-defined parameters, the system allows users to customize the calibration process for atypical series without revising computer code.
- The code assigns default values to all of the user-defined parameters.

Summary

- Although the NEMS is written in Fortran, it can easily communicate with component programs written in R.
- Bayesian techniques allow systematic incorporation of expert judgment into long-term projections.
- By using contributed R packages, EIA is incorporating advanced statistical modeling techniques into the NEMS without programming the new techniques in Fortran.

References

- Bilder, Christopher R. and Loughin, Thomas M. (2015). *Analysis of Categorical Data with R.* CRC Press, New York, pp. 61-140.
- Hernandez, Mari (2013). "Solar Power to the People: The Rise of Rooftop Solar Among the Middle Class." Center for American Progress, October 21, 2013.
- Hilbe, Joseph M. (2011). *Negative Binomial Regression, 2nd edition.* Cambridge University Press, UK.
- Keiser, Richard (2012). "Which are the most profitable regions for solar in the United States?" *PV Tech* article available at http://www.pv-tech.org/guest_blog/21088.
- Rothfield, Emily (2010). "Solar Photovoltaic Installation in California: Understanding the Likelihood of Adoption Given Incentives, Electricity Pricing and Consumer Characteristics." Trinity College of Duke University. Durham, North Carolina.

References (2)

- Harrison, P.J., and Stevens, C. (1971). "A Bayesian approach to short-term forecasting." *Operations Research Quarterly*, **22**, p. 341-362.
- Harrison, P.J., and Stevens, C. (1976). "Bayesian forecasting" (with discussion). *Journal of the Royal Statistical Society*, Series B, **38**, p. 205-247.
- Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*, Springer, New York.
- Welch, G. and Bishop, G. (2006). "An Introduction to the Kalman Filter." Available online at http://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf.
- West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd edition, Springer, New York.

Contact Information

Janice Lent

Janice.lent@eia.gov

202-586-6623

Technical Details

Zero-hurdle models for solar PV

- We first model the binary outcome of a non-zero count being observed for a given zipcode/year combination. For each zipcode z and year t, let

$$y_{0,t,z} = \begin{cases} 0, & \text{if no solar PV installations occur;} \\ 1, & \text{if any solar PV installations occur.} \end{cases}$$

- To model $y_{0,t,z}$, we may use a logit or a probit model.
  - Logit model assumes that the error terms follow a logistic distribution.
  - Probit model assumes that the error terms follow a Gaussian (normal) distribution.

- For the solar PV data, the logit model provided a better fit.

Zero-hurdle models (2)

- Conditional on observing one or more PV installations in a zipcode/year cell, we may use a Poisson or negative binomial distribution to model the actual count of PV installations within the cell.
- Because the solar PV installation counts were over-dispersed for the Poisson model, we used a zero-truncated negative binomial model.
- For each zip code z and year t, let

$$y_{1,t,z} = \text{number of PV installations in year } t \text{ and zipcode } z \mid y_{0,t,z} = 1.$$

Projection method

- We use projected covariates from various NEMS modules to update the hurdle model covariates.
- Because the covariates are projected at the Census Division level, we assume constant multiplicative change factors across zipcodes within each Census Division.
- Apply the hurdle model coefficients to the projected covariate data at the zipcode level.
- Aggregate the zipcode level counts to calculate installations and penetration rates for larger geographic areas.

Components of the multi-process DLM for HGL prices

- Let $k = 4$ be the total number of component models in the multi-process DLM for HGL prices.
- We formulate $k - 1$ component models based on projected values of the Brent crude oil prices (B) and the Henry Hub (H) natural gas prices.
- Let

$$\boldsymbol{y}_t = [y_{C2,t} \quad y_{C3,t}]$$

be the vector of natural logarithms of ethane (C2) and propane (C3) prices in year t.

Components of the multi-process DLM (2)

- We compute initial regression coefficients

$$\boldsymbol{\vartheta}_0 = \begin{bmatrix} \vartheta_{C2,0,0} & \vartheta_{C3,0,0} \\ \vartheta_{C2,B,0} & \vartheta_{C3,B,0} \\ \vartheta_{C2,H,0} & \vartheta_{C3,H,0} \end{bmatrix}$$

  for the intercept terms and the natural logarithms of the Brent and Henry Hub prices, respectively, based on historical data.

- Let $\bar{p}_{H,m,t}$ and $\bar{p}_{B,m,t}$ be the lagged m-year moving averages of the natural logarithms of the projected Henry Hub and Brent prices, respectively, for year t.

Updating DLM components

- For $t > 0$ and for $j = 1, \ldots, k-1$, we set the Henry Hub parameter $\vartheta_{C2,H,t,j}$ for the ethane component model as

$$\vartheta_{C2,H,t,j} = \vartheta_{C2,H,t-1,j} \left[\frac{\left(\frac{\bar{p}_{B,m,t}}{\bar{p}_{B,m,t-1}} \right)}{\left(\frac{\bar{p}_{H,m,t}}{\bar{p}_{H,m,t-1}} \right)} \right]^{\delta_{C2,j}},$$

  where, for $X \in \{1,2\}$, $\delta_{C2,j} \in [0,1]$.

- We define $\vartheta_{C3,H,t,j}$ similarly for propane.

Updating DLM components (2)

- Larger values of $\delta_{C2,j}$ and $\delta_{C3,j}$ may be chosen to reflect the expectation of more pronounced "re-coupling" of price series in response to differences in the rates of change in the Brent and Henry Hub prices.
- To maintain a constant combined effect of the Henry Hub and Brent prices in the component model coefficients, for $t > 0$ and $j = 1, \ldots, k-1$ we set

$$\vartheta_{C2,B,t,j} = \vartheta_{C2,B,0,j} + \vartheta_{C2,H,0,j} - \vartheta_{C2,H,t,j},$$

and

$$\vartheta_{C3,B,t,j} = \vartheta_{C3,B,0,j} + \vartheta_{C3,H,0,j} - \vartheta_{C3,H,t,j}.$$

Weighting the component models

- The choice of component weights reflects the following expectations:
  - Model component 4, whose coefficients are based on historical data, will less accurately reflect market conditions as the projection horizon increases.
  - Of the δ-based model components, we expect either model component 2 (the component with the most moderate δ value) or model component 4 (component with the highest δ value) will most accurately reflect market conditions in the later projection years.

Linear splicing of time series for calibration

- We splice two time series X_t and Y_t, where X_t runs from time $t = 0$ to time $t = t_s$, and Y_t runs from time $t = 0$ to time $t = t_n$, where $t_n > t_s$.
- Splicing preserves the series X_t entirely and modifies the series Y_t to avoid a sharp increase or decrease in the spliced series between time t_s and time $t_s + 1$.
- To perform linear splicing, we choose a "splicing year" t_p, where $t_s < t_p < t_n$.
- For $t = t_s, t_s + 1, \ldots, t_p$, let

$$\hat{Y}_t = X_{t_s} + \left(\frac{t - t_s}{t_p - t_s}\right)\left(Y_{t_p} - X_{t_s}\right).$$

Linear splicing of time series (2)

- The series $\hat{Y}_{t_s}$ increases (or decreases) linearly between the years t_s and t_p.
- The spliced series is

$$Z_t = \begin{cases} X_t, & 0 \le t \le t_s; \\ \hat{Y}_t, & t_s \le t \le t_p; \\ Y_t, & t_p \le t \le t_n. \end{cases}$$

- In our application, the splicing year t_p is based on a year-to-year change tolerance.

Calibration Method 1

- We use Holt-Winters filtering to fit a linear trend series $X_t{}'$ to the short-term series, and we extend the trend series $X_t{}'$ out to the chosen splicing point t_p.
- We check to see which of these two values is smaller:

$$D_s = |Y_s - X_s| \qquad \text{or} \qquad D_s' = |Y_s - X_s{}'|.$$

- If $D_s' < D_s$, we perform *Option 1A*: Linearly splice X_t to $X_t{}'$, then linearly splice $X_t{}'$ to Y_t.
- If $D_s < D_s{}'$, we perform *Option 1B*: Linearly splice X_t to Y_t.
- Method 1 makes no adjustment to the long-term series after the splicing year t_p.

Calibration Method 2

- We fit a linear trend series $X_t{}'$ to the short-term series, as in Option 1.
- We linearly splice X_t to $X_t{}'$ (i.e., we splice the short-term series to its trend series).
- By applying a ratio adjustment, we adjust the level of Y_t to agree with the level of $X_t{}'$ at the point t_p.
- Method 2 adjusts the level of the long-term series to agree with the trend of the short-term series.
- Method 2 is preferred when the level of the original estimated long-term series is uncertain or subject to bias.