Empirical prediction intervals improve energy forecasting

Lynn H. Kaack[a,1], Jay Apt[a], M. Granger Morgan[a], and Patrick McSharry,[b,c]

[a]Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213; [b]Smith School of Enterprise and the Environment, Oxford University, Oxford OX1 3QY, United Kingdom; and [c]Information and Communication Technology (ICT) Center of Excellence, Carnegie Mellon University, Kigali, Rwanda

Edited by B. L. Turner, Arizona State University, Tempe, AZ, and approved July 11, 2017 (received for review December 8, 2016)

Hundreds of organizations and analysts use energy projections, such as those contained in the US Energy Information Administration (EIA)'s Annual Energy Outlook (AEO), for investment and policy decisions. Retrospective analyses of past AEO projections have shown that observed values can differ from the projection by several hundred percent, and thus a thorough treatment of uncertainty is essential. We evaluate the out-of-sample forecasting performance of several empirical density forecasting methods, using the continuous ranked probability score (CRPS). The analysis confirms that a Gaussian density, estimated on past forecasting errors, gives comparatively accurate uncertainty estimates over a variety of energy quantities in the AEO, in particular outperforming scenario projections provided in the AEO. We report probabilistic uncertainties for 18 core quantities of the AEO 2016 projections. Our work frames how to produce, evaluate, and rank probabilistic forecasts in this setting. We propose a log transformation of forecast errors for price projections and a modified nonparametric empirical density forecasting method. Our findings give guidance on how to evaluate and communicate uncertainty in future energy outlooks.

forecast uncertainty | density forecasts | scenarios | continuous ranked probability score | fan chart

Projections of quantities such as electricity and fuel demands, commodity prices, and specific energy consumption and production rates are widely used to inform private and public investment decisions, long-term strategies, and policy analysis (1–3). Policy analysts and decision makers often use modeled projections as forecasts with little or no discussion about the associated uncertainty (2, 4, 5). [Energy outlooks are often referred to as projections because they refrain from incorporating future policy changes into the reference scenario. In contrast, the term forecast denotes a best estimate allowing for all changes of the state of the world (6). While we are aware of this difference, our analysis treats the reference scenario as the best estimate forecast. We use the terms forecast and projection interchangeably.] Here we are concerned with national-scale forecasts in the energy industry that span a range from years to decades. Two of the most influential sets of energy projections are those of the US Energy Information Administration (EIA) and the International Energy Agency (IEA), complemented by those made by private oil and gas companies, such as Shell, ExxonMobil, and Statoil. When assessed retrospectively, such energy projections have sometimes shown very large deviations from the realized values (7–9). Providing information on the likely uncertainty associated with such projections would help individuals and organizations use them in a more informed manner (10–12).

All of the energy outlooks mentioned above provide point projections without a probabilistic treatment of uncertainty. Often, point forecasts are labeled as a "reference scenario" and are accompanied by alternative scenarios. While scenarios may be used to bound a range of possible outcomes, they can easily be misinterpreted (13) and are typically not intended to reflect any treatment of probability. The fact that most projections in the energy space do not report probability distributions around predicted values, or an expected variance, is a problem that has been frequently noted in the literature (13–17). Shlyakhter et al. (14) criticize the EIA for not treating uncertainty in the Annual Energy Outlook (AEO). Density forecasting is increasingly becoming the standard (16, 18) in a variety of disciplines ranging from forecasts of inflation rates (19–21), financial risk management, and trading operations (22, 23) to demographics (24), peak electricity demand (25), and wind power generation (26, 27). There are a number of procedures for probabilistic forecasting (22). Most of these methods take an integrated approach to forecast the whole distribution, including the best estimate. The empirical methods we use here instead allow analysts or forecast users to attach an uncertainty distribution to a preexisting point forecast.

The importance of density forecast evaluation has been discussed by several authors (17, 28–30). When methods are chosen to generate probabilistic energy forecasts, such evaluation is often omitted. Our work is a step toward making energy density forecasting more feasible and robust by framing how to evaluate a probabilistic forecast in this setting.

Choosing a Density Forecasting Method. We compare different methods by testing how accurately they estimate the uncertainty of data that were not used to train the methods.

We argue that if a forecaster is choosing between different methods, this should be the central criterion, even though others such as usability and ease of explanation might also be relevant. Adopting a frequentist's approach, we view a future observation as a random event around the given forecast. A density prediction is best if it equals the probability density function (PDF) from which this future observation is drawn.

Density forecasts are evaluated by their calibration and their sharpness subject to calibration (29). By sharpness we mean that narrower PDFs are preferable. Calibration, as a core concept of

Significance

While many forecasters are moving toward generating probabilistic predictions, energy forecasts typically still consist of point projections and scenarios without associated probabilities. Empirical density forecasting methods provide a probabilistic amendment to existing point forecasts. Here we lay the groundwork for evaluating the performance of these methods in the data-scarce setting of long-term forecasts. Results can give policy analysts and other users confidence in estimating forecast uncertainties with empirical methods.

Author contributions: L.H.K., J.A., and M.G.M. designed research; L.H.K., J.A., M.G.M., and P.M. performed research; L.H.K. and P.M. contributed new reagents/analytic tools; L.H.K. analyzed data; and L.H.K., J.A., M.G.M., and P.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: kaack@cmu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1619938114/-/DCSupplemental.

Table 1. Empirical density forecasting methods compared

Method	Parametric	Based on	Median centered
NP_1: nonparametric EPI	No	Forecast errors	No
NP_2: nonparametric centered EPI	No	Forecast errors	Yes
G_1: Gaussian distribution	Yes	Forecast errors	Yes
G_2: Gaussian distribution	Yes	Historical deviations	Yes

Details can be found in *Materials and Methods*.

forecast evaluation, refers to the predictive density representing correctly the true PDF of the observation. Measuring calibration requires the availability of unknown observations. This can be simulated by using an early portion of the time series to train the density prediction and using later actual values as the test observations. This procedure is referred to as out-of-sample forecast evaluation. Dividing the data into these two sets requires a long enough record of historical data and forecasts to draw statistically significant conclusions. While the AEO sample size is small, we see no viable alternative to this procedure and find that even small sample results can provide useful insights.

As it is a measure of both calibration and sharpness, we use the continuous ranked probability score (CRPS) (30–32) to compare density forecasts. For point forecast evaluation we work with the average prediction error, here the mean absolute percentage error (MAPE), and the transformed mean absolute logarithmic error (MALE) for prices (*Materials and Methods*).

Empirical Density Prediction Methods. We compare four different data-driven parametric and nonparametric estimates of forecast uncertainty in the form of PDFs (Table 1 and *Materials and Methods*). A simple method of empirical prediction intervals (EPIs), first published by Williams and Goodman (33), uses the distribution of past forecast errors to create a probability density forecast around an existing point forecast. It relies on the assumption that past errors are a good estimator of the forecaster's current ability to predict the future. EPIs are an established approach and have been used in a number of fields such as meteorology (34), including the creation of the classic "cone of uncertainty" now routinely produced for likely hurricane tracks (35), future commodity prices (36), and the values of macroeconomic variables such as inflation (20). There is a continuing interest in the method from researchers in applied mathematics and statistics (18, 37, 38). We introduce a second nonparametric EPI, which is a modification of Williams and Goodman's EPI, with a centered error distribution. For a third, parametric, prediction method we use the forecasting errors to estimate a Gaussian density forecast. A parametric PDF has the advantage of greater ease of use. We use the volatility of the time series of historical values to inform a fourth probabilistic forecast, which is valuable in cases where the forecasting record is short.

We apply the four different methods to 18 quantities in EIA's AEO (39), which are chosen based on EIA's Retrospective Review (40) (*Materials and Methods*). The AEO forecasting record spans more than 30 years. Unfortunately, in the context of forecast evaluation a sample size of ∼30 data points is very small. In addition, because of modifications that EIA makes to its models, and changes in technology, market conditions, and regulations, errors are not likely to be stationary. Because stationarity of past forecasting errors is an essential requirement for good performance of EPIs (38), we test the extent to which PDFs estimated using this procedure provide robust probabilistic forecasts. Previous work has analyzed the forecast errors of EIA's AEO (1–3, 7, 41, 42) and the projections by the IEA (8). Generally, authors have focused on a mean percentage error and directional consistency of errors, also termed bias. Shlyakhter et al. (14) constructed a parametric density forecast with the retrospective errors of AEOs, similar to what we test in this paper. However, they did not assess the calibration of their prediction intervals.

We begin by evaluating the point forecast performance of the AEO reference case over our test range of AEO 2003–2014. Using the same out-of-sample AEOs and historical observations, we then compare the calibration and sharpness of the four different density forecasts. The prediction intervals are also compared with the scenarios published in the AEO. We find that over the test range a normal distribution based on past forecasting errors clearly outperformed uncertainties based on the scenarios in the AEO. This conclusion is for the diverse set of all quantities, but depending upon the quantity, in some cases other methods showed better results. We conclude the paper with a comparative discussion of the methods and their applicability to energy forecasting.

Results

We evaluate the predictive performance of four uncertainty estimation methods (Table 1) over the test range of AEO 2003–2014 and observations of 2002–2015, using 1985–2002 as the training range. The test range excludes AEO 2009, which did not provide scenarios for the updated reference case. We determine the number of quantities for which a method performed best. We find that Gaussian densities informed by retrospective errors (G_1) or based on the variability of the historical values (G_2) performed best for the most quantities. The original nonparametric method, as in ref. 33 (NP_1), performed best in very few cases. The centered nonparametric distribution (NP_2), which gives the largest weight to the AEO reference case projection instead of the bias, performed better over the test range than NP_1. The respective best empirical uncertainty estimation methods had significantly better calibration than methods based on the AEO scenarios with 95% confidence. In fact, G_1 significantly outperformed the scenarios for all quantities and provided a valid general approach to estimate the uncertainty in the AEO.

While we have performed analysis for 18 quantities forecasted in the AEO, we use 2 of the quantities, natural gas wellhead price in nominal dollars per 1,000 cubic feet (hereafter natural gas price) and total electricity sales in billion kilowatt hours

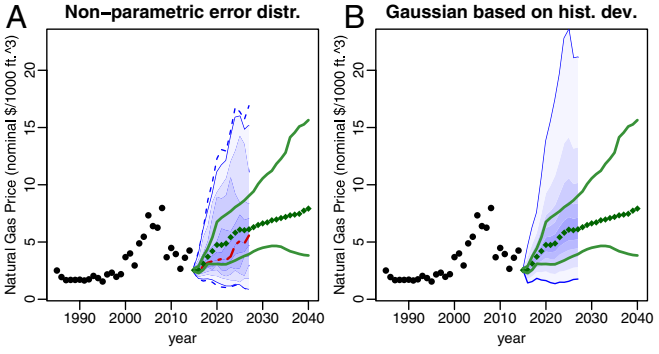


Fig. 1. Density forecasts for natural gas prices in nominal dollars. (*A*) Nonparametric EPI based on forecast errors (NP_1). (*B*) Gaussian density forecast based on the variability of historical values (G_2), which tested to be the better estimate. Historical values are indicated by black circles, the AEO 2016 reference case by green diamonds, and the density forecast by blue shaded areas. The different shades correspond to the percentiles 2, 10, 20, 30, ..., 80, 90, 98. The outermost dashed lines report the minimum and maximum value of the error samples. AEO 2016 envelope scenarios are in green. Note that in *A* the median of the predictive distribution (dashed red line) does not coincide with the reference case.

SUSTAINABILITY SCIENCE

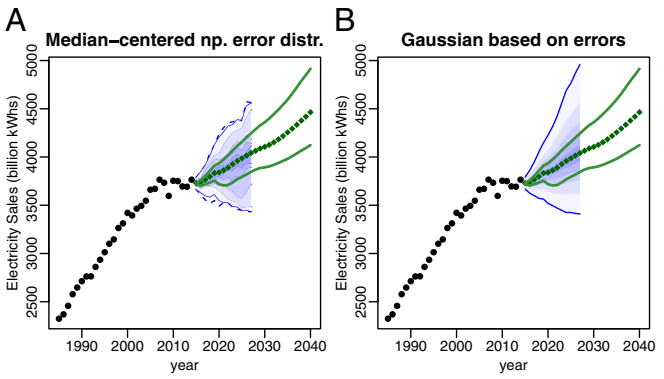


Fig. 2. Density forecasts for electricity sales based on AEO 2016. (*A*) For the median-centered nonparametric EPI (NP_2), the median or bias now coincides with the AEO reference case. (*B*) The Gaussian density forecast based on the SD of the errors (G_1) was the best forecast over the test range. The envelope scenarios are narrower.

(hereafter electricity sales), for illustration purposes (Figs. 1 and 2). Results for all 18 quantities can be found in *SI Appendix*.

Error Metric and Transformation for Price Quantities. All forecast evaluation scores are computed on the basis of the deviations of the forecasts $\hat{y}$ from historical values y, referred to as error. We found it useful to work with the percentage error, or relative error, $\epsilon_{rel} = \frac{\hat{y}-y}{y} = \frac{\hat{y}}{y} - 1$. Percentage errors allow us to compare different quantities and they are independent of changes in the currency value. We can conduct the analysis in a similar way with absolute errors. Since the error distributions of price quantities are asymmetric, as prices are typically log-normally distributed (43), we modify the error for price quantities. Drawing an analogy to logarithmic returns, a concept from financial theory, we modify ϵ_{rel} to yield the logarithmic error $\epsilon_{log} = \ln(1 + \epsilon_{rel}) = \ln\left(\frac{\hat{y}}{y}\right) = \ln\hat{y} - \ln y$. For prices we compute the comparative statistics and additional transformations, such as centering of the PDF, in ϵ_{log} (*SI Appendix*).

The structure of the relative errors as a function of forecast year and forecast horizon is shown in Fig. 3. The horizon H refers to the number of time steps, or years, into the future that the forecast is made. Uncertainty increases with H. AEO projections reflect uncertainty in past values; e.g., for AEO 2016 we therefore refer to 2015 as $H = 0$ and 2016 as $H = 1$.

Retrospective Analysis Can Inform Density Forecasts. We illustrate examples of the four probabilistic forecasting methods listed in Table 1. Figs. 1 and 2 compare the nonparametric methods to the methods that performed better for the two example quantities, that is, the two Gaussian predictions.

A nonparametric distribution of the errors (NP_1) results in the EPI shown in Fig. 1*A*. Here the median of the errors is not exactly zero, which is often referred to as bias. We see that this results in a second point forecast or a best estimate forecast that is not equal to the reference case scenario. If we can assume that the forecasting errors are stationary, then past and future errors follow the same PDF, and this bias should yield a better point forecast than the reference case. However, we found this is not the case for most quantities.

Modifying the nonparametric distribution in such way that it places the greatest weight on the AEO reference case projection is one approach to combat this problem (NP_2). This centered EPI for electricity sales is shown in Fig. 2*A*. In the percentage-error space, we center by subtracting the median error m_{rel} from all errors in the distribution $\epsilon_{rel,ctr} = \epsilon_{rel} - m_{rel}$. For the price quantities, we transform the distribution in log-error space. We define the log median $m_{log} = \text{median}(\epsilon_{log}) = \ln(1 + m_{rel})$. The centered log errors are then $\epsilon_{log,ctr} = \epsilon_{log} - m_{log} = \ln\left(\frac{1+\epsilon_{rel}}{1+m_{rel}}\right)$ (*SI Appendix*).

These two nonparametric estimations are compared with two parametric distributions, Gaussians with a mean of zero and the variance of the errors (G_1) (Fig. 2*B*) and with the variance of historical values (G_2) (Fig. 1*B*). When modeling normality, we implicitly make assumptions about the nature of the errors. Extreme errors, which can have large consequences for decision making, occur frequently in energy forecasting (14). A Gaussian PDF may not do an adequate job of representing heavier tails and might underestimate the probability of extreme events. However, a parametric distribution will generate longer tails than a nonparametric error PDF. Regarding usability, the simplicity of a two-parameter specification prevails over nonparametric distributions. A discussion of normality and correlation in the errors is provided in *SI Appendix*.

Past Bias in the AEO Does Not Predict Future Bias. Recently, electricity sales have been flat. Can a forecast be better than a constant prediction using the last observation, i.e., persistence? We can assess the point forecasting skill of the AEO reference case projections by comparing them with benchmark forecasts such as persistence or simple linear regression. To compare different point forecasts, we evaluate the MAPE and the MALE for prices. MAPE and MALE are defined as the sum over the absolute value of all observed errors for a given horizon (*Materials and Methods*). A larger MAPE/MALE indicates that the forecast has performed worse over the test range 2003–2014 (Fig. 4).

We find that persistence performed surprisingly well over the test range of the last decade, outperforming the AEO for 10 of the 18 quantities. This is due to the fact that the recent decade has seen trend changes that are conducive to persistence forecasts. If the length of the fitted window is optimized for the test range, a simple linear regression significantly outperforms the reference case for eight quantities with 95% confidence. Point forecast comparison of the AEO reference case with the median of the errors reveals that correcting for the bias is not a good strategy in most cases. The AEO reference case was a better point forecast than the bias for most of the quantities over the test range, except for coal production and residential energy consumption. We therefore anticipate that centering the nonparametric uncertainty (NP_2) is advised for most quantities except those.

Gaussian Density Forecasts Often Perform Well. Scoring rules, or scores, provide a means for comparing the performance of different probabilistic forecasts. We use the CRPS, which is a strictly

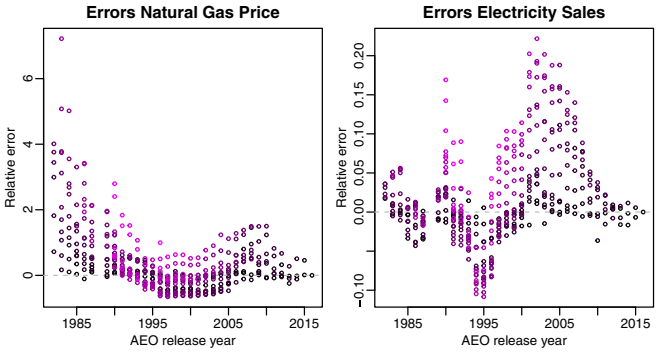


Fig. 3. Forecast errors by AEO release year. Different colors correspond to forecast horizons ranging from $H = 0$ in black to $H = 21$ in purple. All forecast errors are untransformed. Note the different scale. No AEO was released for 1988.

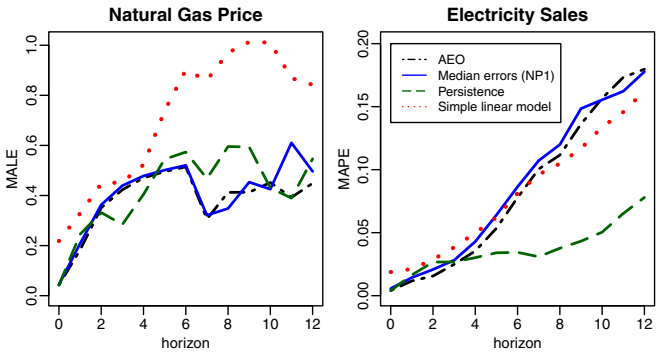


Fig. 4. The mean absolute percentage or log error (MAPE/MALE) for the test range 2003–2014. We see that for natural gas prices (in nominal dollars), the median of NP_1 performs similarly to the AEO reference case. For electricity sales, the reference case outperforms the median for nearly every horizon. For the test range, a persistence forecast has clearly been the best forecast for electricity sales, which have recently experienced near zero growth.

proper score in this case (31). It assigns value not only to the predicted probability of an observation but also to the distance of a predicted probability mass from an observation. It is therefore relatively robust to specific functional forms of the density forecasts (30) and allows for comparison with point and ensemble forecasts (31, 32) (*Materials and Methods*).

The results of the average CRPS over the test range for each horizon in units of relative or log error are illustrated in Fig. 5. A standalone value of the CRPS is not meaningful; it serves to provide a comparison between different methods. As the CRPS reduces to the MAPE/MALE for a point forecast, it is informative to compare the results to the MAPE/MALE of the AEO reference case. In Fig. 5, we find that the scenarios (S) only marginally improve the prediction with respect to the point forecast. In addition, we see that for the natural gas price, NP_1 is larger than the MALE due to poor point forecast performance of the EPI's median.

To find the best density prediction method, we normalize the CRPS of each method by the CRPS of the scenario ensemble (S) for every horizon (Fig. 6). For every quantity, we then average over a core range of horizons $H = 2$ to $H = 9$ and rank these aggregated scores. The method with the lowest average rank is considered the best density over the test range for a given quantity. We find that the results barely change if more horizons, modifications to the test range, or an alternative ranking method are considered (*SI Appendix*).

The ranking of all quantities shows that the two Gaussian methods perform well for most quantities (Fig. 7). G_1 counts as the best method for 9 of the 18 quantities and G_2 for 3 quantities. The performance of G_2 is, however, often similar to that of G_1 and it is second best for 8 quantities. The fact that these parametric methods performed well over the test range is convenient, because there are standard ways to use a normal distribution as a model input. Besides these parametric methods, also NP_2 performed well. As expected, in the two cases of coal production and residential energy consumption, including the bias with NP_1 seemed the best approach over the test range. In the following section, we analyze whether the empirical methods performed significantly better than uncertainty estimates based on the scenarios.

AEO Scenario Ranges Are Narrower Than Observed Uncertainties. Every AEO includes a number of scenarios, intended as sensitivity studies on the reference case under a small number of varied input assumptions. No value is assigned to the probability that a future outcome will lie within the scenario range. The CRPS allows for comparison of a density forecast with an ensemble forecast. It assigns every discrete scenario an equal point probability mass (S). Because of the varying number of scenarios in the AEO, we make a simplification and consider only the reference case and the high- and low-envelope scenarios, which do not correspond to a specific scenario in the AEO (*Materials and Methods*). In addition, we discuss a Gaussian distribution (SP_1) and a uniform distribution (SP_2) based on the envelope scenarios.

The CRPS scores normalized by the score of S are shown in Fig. 6. Fig. 6 also includes the scores for the sensitivity cases SP_1 and SP_2. A normalized CRPS of an empirical method that is <1.0 indicates an improvement over uncertainties based on the scenarios (S). We can find at least one density forecasting method for every quantity, which on average over the core horizons performed better than the scenarios. In addition, we conduct a hypothesis test if we can reject that either S or SP_1 was the better probabilistic forecast over the test range. We find that the best-ranked empirical method for a respective quantity was significantly better than both S and SP_1 with 95% confidence. In fact, NP_2, G_1, and G_2 all show significant improvements (Fig. 7). These results are likely due to the fact that over the test range on average the scenario range of all AEO quantities covered only 14% of the actual values (*SI Appendix*). The width between the highest and the lowest scenario, however, changes greatly from one AEO to another and is somewhat correlated to the number of scenarios published.

Discussion and Conclusion

This analysis showed that empirical density prediction methods, based on forecasting errors or historical deviations, provide valuable approaches for including an estimate of uncertainty with a forecast. There are empirical methods available for estimating the uncertainty around the AEO reference case, which have proved to be significantly more accurate over the past decade than the scenarios of the AEO. We find that a Gaussian distribution based on past errors (G_1) offers a method with convincing ease of use and good performance over the different quantities (Fig. 7). We therefore recommend that the EIA and others producing energy forecasts include the SD of forecast errors in their retrospective reports. We supply the values for AEO 2016 in *SI Appendix*. A nonparametric distribution of the observed forecast errors was the better density forecast only in a few cases, confirming that focusing on representing the exact error distribution does not need to provide the better out-of-sample forecast. Point forecast evaluation illuminated that EIA's forecast bias is in most cases not consistent and that using a bias-corrected reference case does typically not lead to the better forecast.

As both the forecasting process and the energy system can be nonstationary, there is no way to be sure that our results will be applicable to future data. However, the way we evaluated and

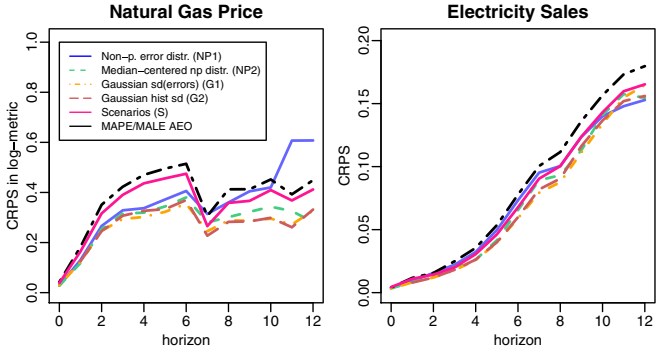


Fig. 5. The CRPS for the test range 2003–2014. A lower CRPS corresponds to a better density or ensemble forecast.

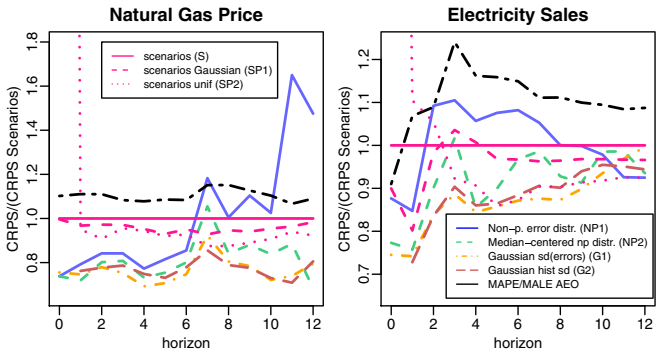


Fig. 6. Relative improvement of the methods with respect to the envelope scenarios for the test range 2003–2014. Values are plotted as fraction of the CPRS of the scenario ensemble (S). A normalized CRPS lower than 1.0 corresponds to a better density forecast. SP_1 corresponds to a normal distribution with the scenario range as 1 SD, and SP_2 is a uniform PDF between the envelope scenarios.

chose a method is a robust procedure. Hence, in the absence of other insights we recommend using one of the Gaussian distributions.

Despite the advantages of probabilistic forecasts, scenarios convey important information about the workings of energy predictions and allow users to better understand and compare the assumptions. We emphasize that the combined use of a density forecast and scenarios would be a fruitful approach to describe the uncertainty of a forecast. Empirical density forecasts are easily reproducible, but other probabilistic methods such as a quantile forecasting could also advance energy projections.

Materials and Methods

See *SI Appendix* for a detailed description of the materials and methods used.

Data. The dataset consists of AEOs 1982–2016 and historical values from 1985 to 2015. Historical data were taken from the EIA Retrospective Review (40) and the AEOs (39), and conversions were applied where necessary. All data are publicly available on the EIA website. Refer to *SI Appendix: Data Description* for more detail. The data analysis was performed in R (44).

List of Methods.

Point forecasting methods.

AEO reference case. We treat the AEO reference case as a point forecast. The reference case is a projection of the current state of laws and regulations and does not represent a best estimate forecast. Also the EIA chooses the reference case as a best estimate when determining projection errors (40).

Median errors (NP_1). The median of the EPI with a nonparametric distribution of the errors (NP_1) is computed as the reference case adjusted by the median of past forecasting errors.

Persistence. Persistence refers to a constant forecast equal to the last observation. Here, we use the forecasted value at $H = 0$ as the last observation, since on the AEO release date this is the closest approximation to the actual value.

Simple linear model. This benchmark is a simple linear regression with time as the predictor. The quantity is regressed over a moving window of the last seven historical observations. This size of window is the optimum for the test range.

Density forecasting methods.

NP_1. This method is an EPI with a nonparametric distribution of the forecasting errors and a median different from the reference case. This method was originally published by ref. 33.

NP_2. This method is an EPI with a nonparametric error distribution, which is centered such that the median and $\epsilon = 0$ align. This results in the AEO reference case being the best estimate forecast.

G_1. This method is a Gaussian distribution with the SD of the past errors and a mean and median of $\epsilon = 0$.

G_2. This method is a Gaussian distribution with a SD based on a sample of all relative deviations between two historical data points which are H steps apart. Mean and median are $\epsilon = 0$.

S. This ensemble forecast consists of the reference case and the highest and lowest scenario projections in every year. These correspond to the envelope of all scenarios by using only the highest and lowest projected values.

SP. Two parametric density predictions are based on the envelope scenarios in the AEO. We chose a Gaussian distribution with the distance to the farthest scenario as 1 SD (SP_1) and a uniform distribution between the envelope scenarios (SP_2).

MAPE. The MAPE is a measure for point forecast performance. This becomes the MALE in the case of price forecasts with log errors. They are defined as

$$MAPE_H = \frac{1}{n_H}\sum_{t=1}^{n_H}|\xi_{rel,H,t}| = \frac{1}{n_H}\sum_{t=1}^{n_H}\left|\frac{\hat{y}_{H,t}-y_{H,t}}{y_{H,t}}\right|, \quad [1]$$

and $MALE_H = \frac{1}{n_H}\sum_{t=1}^{n_H}|\ln\hat{y}_{H,t} - \ln y_{H,t}|$, where there are n_H errors for a particular horizon H. $\hat{y}$ refers to the forecast, while y is the actual observation.

CRPS. The CPRS for every horizon, as we use it in this paper, is defined as

$$CRPS_H(F,\epsilon) = \frac{1}{n_H}\sum_{t=1}^{n_H}\int_{-\infty}^{\infty}(F_t(\epsilon_t) - I(\epsilon_t \geq \xi_t))^2 d\epsilon_t \quad [2]$$

similar to ref. 31. ϵ_t is a point of the predictive error distribution, while ξ_t is the forecast error of the observation. The CRPS compares the cumulative distribution function (CDF) of the density forecast with the CDF of an observation, a step function $I(\epsilon_t \geq \xi_t)$. We compute the score in the respective error metric. The CRPS for a nonparametric CDF is computed like the CRPS for an ensemble forecast of discrete scenarios (32). For ensemble forecasts, the CRPS can also be written as $CRPS_H(F,\epsilon) = \frac{1}{n_H}\sum_{t=1}^{n_H}\left[E_F\left|\epsilon_t - \xi_t\right| - \frac{1}{2}E_F\left|\epsilon_t - \epsilon'_t\right|\right]$ (31). In our case, the $CRPS_H$ reduces to the $MAPE_H$ for a point forecast. In this case we have a single $\epsilon_t = 0$, resulting in $E_F\left|\epsilon_t - \xi_t\right| = |\xi_t|$ and $E_F\left|\epsilon_t - \epsilon'_t\right| = 0$. The CRPS is a strictly proper score here (31), which means that the expected score is maximized if the observation is drawn from the predictive distribution and this maximum is unique. The CRPS has different scales for different quantities or error measures, which is why we normalize the $CRPS_H$ by the $CRPS_{S,H}$ of the scenario ensemble.

Improvement Testing. We perform a bootstrap on the single CRPS results in a horizon sample, which then is used to compute the $CRPS_H$ and the

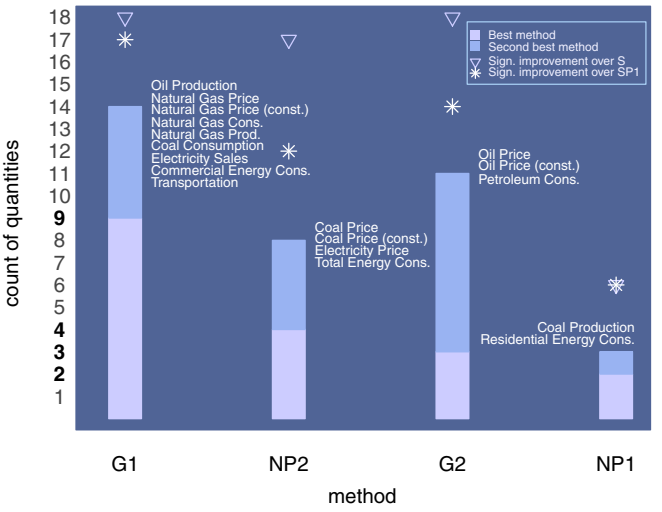


Fig. 7. Graphical summary of the evaluation results. The methods are ordered by the number of quantities they perform best for (listed in white). The Gaussian based on errors (G_1) performs best or second best for 14 of 18 and showed significant improvement over the scenarios for almost all quantities. Improvement is more likely over S than over SP_1. The nonparametric biased EPI (NP_1) performs worse than the nonparametric centered EPI (NP_2) and the Gaussian based on historical deviations (G_2).

aggregated CRPS average for the ranking. For each of the four methods, we determine the portion of resampled results that indicates that S or SP_1 is the better forecast. If this portion is smaller than 0.05, we speak of the method as being a significant improvement over the scenarios.

Sensitivity Analysis On the Ranking Results. To test the sensitivity of the ranking, we varied the default assumptions. Instead of first averaging the normalized CRPS and then ranking that result, we alternatively first ranked the $CRPS_H$ and then averaged over the horizons. We also averaged over the full range of horizons $H = 1$ to $H = 12$ instead of the core range that included large H with small sample sizes. In addition, we included AEO 2009 in the test range. The respective best methods did not change with these variations. For some quantities, the performances of the best and second-best methods were very similar to each other. This resulted in a sensitivity regarding a change in the test range for three quantities.

ACKNOWLEDGMENTS. We thank Evan D. Sherwin, Inês L. Azevedo, Cosma R. Shalizi, Alexander L. Davis, Stephen E. Fienberg, and Max Henrion for their advice and assistance. Evan D. Sherwin led the data collection and adjustments. We thank the EIA for hosting a presentation and discussion about this work, in particular Faouzi Aloulou, David Daniels, and John Staub. This work was supported by the Electric Power Research Institute and by the Center for Climate and Energy Decision Making through a cooperative agreement between the National Science Foundation and Carnegie Mellon University (SES-0949710).

1. Winebrake JJ, Sakva D (2006) An evaluation of errors in US energy forecasts: 1982–2003. *Energy Policy* 34:3475–3483.
2. Wara M, Cullenward D, Teitelbaum R (2015) Peak electricity and the clean power plan. *Electr J* 28:18–27.
3. Gilbert AQ, Sovacool BK (2016) Looking the wrong way: Bias, renewable electricity, and energy modelling in the United States. *Energy* 94:533–541.
4. Neuhauser A (2015) Wasted energy. *US News World Rep*. Available at https://www.usnews.com/news/articles/2015/05/28/wasted-energy-the-pitfalls-of-the-eias-policy-neutral-approach. Accessed July 23, 2017.
5. Harvey C (2016) How we get energy is changing fast—and it's sparking a huge fight over forecasting the future. *Wash Post*. Available at https://www.washingtonpost.com/news/energy-environment/wp/2016/05/13/how-we-get-energy-is-changing-rapidly-and-its-sparking-a-huge-fight-over-forecasting-the-future/?utm_term=.987c4550ffc8.
6. Intergovernmental Panel on Climate change (2015) Definition of terms used within the DDC pages. Available at www.ipcc-data.org/guidelines/pages/definitions.html. Accessed July 23, 2017.
7. Fischer C, Herrnstadt E, Morgenstern R (2009) Understanding errors in EIA projections of energy demand. *Resour Energy Econ* 31:198–209.
8. Linderoth H (2002) Forecast errors in IEA-countries' energy consumption. *Energy Policy* 30:53–61.
9. Smil V (2000) Perils of long-range energy forecasting: Reflections on looking far ahead. *Technol Forecast Soc Change* 65:251–264.
10. Schlaifer R, Raiffa H (1961) *Applied Statistical Decision Theory* (Division of Research, Harvard Business School, Boston).
11. Morgan MG, Henrion M (1990) *Uncertainty : A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis* (Cambridge Univ Press, Cambridge, UK).
12. Fischhoff B, Davis AL (2014) Communicating scientific uncertainty. *Proc Natl Acad Sci USA* 111:13664–13671.
13. Morgan MG, Keith DW (2008) Improving the way we think about projecting future energy use and emissions of carbon dioxide. *Clim Change* 90:189–215.
14. Shlyakhter AI, Kammen DM, Broido CL, Wilson R (1994) Quantifying the credibility of energy projections from trends in past data: The US energy sector. *Energy Policy* 22:119–130.
15. Craig PP, Gadgil A, Koomey JG (2002) What can history teach us? A retrospective examination of long-term energy forecasts for the United States. *Annu Rev Energy Environ* 27:83–118.
16. Gneiting T (2008) Editorial: Probabilistic forecasting. *J R Stat Soc Ser A Stat Soc* 171:319–321.
17. Vahey SP, Wakerly L (2013) Moving towards probability forecasting. *Globalisation and Inflation Dynamics in Asia and the Pacific* (Bank for International Settlements, Basel, Switzerland), BIS Paper 70b. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2248763. Accessed July 23, 2017.
18. Gneiting T, Katzfuss M (2014) Probabilistic forecasting. *Annu Rev Stat Appl* 1:125–151.
19. Diebold FX, Tay AS, Wallis KF (1997) Evaluating density forecasts of inflation: The survey of professional forecasters. *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, eds Engle RF, White H (Oxford Univ Press, Oxford), pp 76–90.
20. Britton E, Fisher P, Whitley J (1998) The inflation report projections: Understanding the fan chart. *Bank Engl Q Bull* 38:30–37.
21. Blix M, Sellin P (1998) Uncertainty bands for inflation forecasts. Available at www.riksbank.se/en/Press-and-published/Published-from-the-Riksbank/Other-reports/Working-Paper-Series/1998/No-65-Uncertainty-Bands-for-Inflation-Forecasts/. Accessed July 23, 2017.
22. Tay AS, Wallis KF (2000) Density forecasting: A survey. *J Forecast* 19:235–254.
23. Linsmeier TJ, Pearson ND (2000) Value at risk. *Financial Analysts J* 56:47–67.
24. Raftery AE, Li N, Ševčíková H, Gerland P, Heilig GK (2012) Bayesian probabilistic population projections for all countries. *Proc Natl Acad Sci USA* 109:13915–13921.
25. McSharry PE, Bouwman S, Bloemhof G (2005) Probabilistic forecasts of the magnitude and timing of peak electricity demand. *IEEE Trans Power Syst* 20:1166–1172.
26. Taylor JW, McSharry PE, Buizza R (2009) Wind power density forecasting using ensemble predictions and time series models. *IEEE Trans Energy Convers* 24:775–782.
27. Pinson P (2013) Wind energy: Forecasting challenges for its operational management. *Stat Sci* 28:564–585.
28. Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts, with applications to financial risk management. *Int Econ Rev* 39:863–883.
29. Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *J R Stat Soc Series B Stat Methodol* 69:243–268.
30. Smith LA, Suckling EB, Thompson EL, Maynard T, Du H (2015) Towards improving the framework for probabilistic forecast evaluation. *Clim Change* 132:31–45.
31. Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 102:359–378.
32. Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast* 15:559–570.
33. Williams WH, Goodman ML (1971) A simple method for the construction of empirical confidence limits for economic forecasts. *J Am Stat Assoc* 66:752–754.
34. Pinson P, Kariniotakis G (2010) Conditional prediction intervals of wind power generation. *IEEE Trans Power Syst* 25:1845–1856.
35. NOAA National Hurricane Center (2016) *National Hurricane Center Forecast Verification.* Available at www.nhc.noaa.gov/verification/verify6.shtml. Accessed July 23, 2017.
36. Isengildina-Massa O, Irwin S, Good DL, Massa L (2011) Empirical confidence intervals for USDA commodity price forecasts. *Appl Econ* 43:3789–3803.
37. Knüppel M (2014) Efficient estimation of forecast uncertainty based on recent forecast errors. *Int J Forecast* 30:257–267.
38. Lee YS, Scholtes S (2014) Empirical prediction intervals revisited. *Int J Forecast* 30:217–234.
39. US Energy Information Administration (2016) *Annual Energy Outlook*. Available at www.eia.gov/forecasts/aeo/. Accessed July 23, 2017.
40. US Energy Information Administration (2015) *Annual Energy Outlook Retrospective Review.* Available at https://www.eia.gov/forecasts/aeo/retrospective/. Accessed July 23, 2017.
41. O'Neill BC, Desai M (2005) Accuracy of past projections of US energy consumption. *Energy Policy* 33:979–993.
42. Auffhammer M (2007) The rationality of EIA forecasts under symmetric and asymmetric loss. *Resour Energy Econ* 29:102–121.
43. Sprenkle CM (1961) Warrant prices as indicators of expectations and preferences. *Yale Econ Essays* 1:179–231.
44. R Core Team (2015) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).

Supporting Information (SI Appendix)

Kaack et al. "Empirical Prediction Intervals Improve Energy Forecasting"

1. Data Description

All data come from the Annual Energy Outlook [1] and the Retrospective Review [2] of the U.S. Energy Information Administration (EIA). The data set consists of AEO 1982-2016, with historical, or actual, values for 1985-2015. Historical values are taken from the EIA's Retrospective Reviews with the exception of 2014 and 2015 which are taken from AEO 2016 [1]. Historical values for 2015 are the $H = 0$ projections from AEO 2016, which might be updated in the following AEO. Any missing values are linearly interpolated.

Before 1988, the AEO was released in the end of each fiscal year and after 1988 in April of the following year. This renaming decision led to the fact that there is no AEO 1988. For reasons of simplicity, we will use naming conventions based on the AEOs released after 1988. Horizons in our data set range from $H = 0$ to $H = 21$. As the collection of historical data is not complete when the forecasts are issued, AEOs include estimates of the year before the release and a forecast of the year of the release. AEO 2000 for example has estimates for 1999 and 2000. We refer to these estimates as forecast horizons $H = 0$ and $H = 1$ respectively. The number of forecasting errors for each horizon varies from $n_{H=0\cdots3} = 31$ to $n_{H=21} = 1$. As sample sizes are decreasing with larger horizons and the variance of errors is dependent on the sample size, we chose a maximum horizon for the analysis of $H_{max} = 12$, where $n_{H=12} = 19$ and $n_{H=13} = 16$.

The AEO projections are based on the National Energy Modeling System (NEMS). The EIA ensures that projections match across its products. For shorter time horizons (up to two years ahead), the EIA arranges that the NEMS outputs are consistent with the forecasts in the Short-Term Energy Outlook (STEO) [3]. The STEO is based on a different forecasting system and contains forecasts as opposed to projections. This does, however, not impact our analysis.

As the AEO projects a large number of quantities, we restrict ourselves to eighteen select quantities of EIA's Retrospective Review [2]. The quantity names used throughout the paper correspond to the following AEO naming conventions:

1. *Oil Price (nominal dollars):* Imported refiner acquisition cost of crude oil in nominal dollars per barrel; also crude oil spot prices, crude oil prices, world oil price
2. *Oil Price (constant dollars):* Imported refiner acquisition cost of crude oil in constant 2013 dollars per barrel; also crude oil spot prices, crude oil prices, world oil price
3. *Petroleum Cons.:* Total petroleum consumption in million barrels per day; also liquid fuels: primary supply, product supplied: total product supplied, liquid fuel consumption: total, refined petroleum products supplied: total, petroleum product supplied
4. *Oil Production:* Domestic crude oil production in million barrels per day; also liquid fuels: crude oil: domestic production, domestic crude production, production: crude oil, petroleum production: crude oil
5. *Natural Gas Price (nom.):* Natural gas wellhead prices in nominal dollars per thousand cubic feet; also Henry Hub spot price, average lower 48 wellhead price
6. *Natural Gas Price (const.):* Natural gas wellhead prices in constant 2013 dollars per thousand cubic feet; also Henry Hub spot price, average lower 48 wellhead price
7. *Natural Gas Consumption:* Total natural gas consumption in trillion cubic feet; also natural gas: use by sector: total, consumption by sector: total
8. *Natural Gas Production:* Natural gas production in trillion cubic feet; also dry gas production
9. *Coal Price (nom.):* Coal prices to electric generating plants in nominal dollars per million Btu; also delivered prices: electric power
10. *Coal Price (const.):* Coal prices to electric generating plants in constant 2013 dollars per million Btu; also delivered prices: electric power
11. *Coal Consumption:* Total coal consumption in million short tons; also coal supply: use by sector: total, consumption by sector: total, total consumption
12. *Coal Production:* Coal production in million short tons, this includes waste coal supplied; also production: total and waste coal supplied, production: total, coal production
13. *Electricity Price:* Average electricity prices in nominal cents per kilowatt-hour; also end-use prices: all sectors average
14. *Electricity Sales:* Total electricity sales in billion kilowatthours; also electricity sales by sector: total, generation by fuel type: total electricity sales
15. *Total Energy Cons.:* Total energy consumption in quadrillion Btu; also energy use: delivered: all sectors: total, delivered energy consumption: all sectors: total, primary energy consumption: total
16. *Residential Energy Cons.:* Total delivered residential energy consumption in quadrillion Btu; also energy use: residential: delivered energy, residential: total
17. *Commercial Energy Cons.:* Total delivered commercial energy consumption in quadrillion Btu; also energy use: commercial: delivered energy, commercial: total
18. *Transportation:* Total delivered transportation energy consumption in quadrillion Btu; also energy use: transport: delivered energy, transportation: total

We excluded total delivered industrial energy consumption, which is a quantity in the Retrospective Review, based on a change in definition by the EIA which we could not correct

for. We are able to generate probabilistic forecasts for total energy related carbon dioxide emissions, but we excluded it from the final analysis due to the shorter forecasting record. The EIA only began to publish carbon dioxide emissions in the AEO 1993.

We analyze each quantity to find the most general approach to creating and evaluating the probabilistic forecasts. We use two of the quantities for illustration purposes in the main article: As prices exhibit a larger degree of volatility than other quantities, we chose to include one price forecast and one other quantity. The **natural gas wellhead price** in nominal dollars per 1000 cubic ft. (hereafter natural gas price) is an important factor for investment decisions. The EIA Retrospective Reviews [2] note the large differences of natural gas price projections and historical values. The Retrospective Review published in 2014 describes that natural gas price predictions influence gas consumption and electricity price forecasts, and recently also coal consumption projections [2]. An example with less volatile historical values are the **total electricity sales** in billion kWhs. The EIA points out the large underestimation of electricity sales in the nineties and the effect on the coal consumption forecasts in its 2008 Retrospective Review [2].

In Fig. S3 we see the historical actual data and the past AEO reference case projections for the two quantities selected. This figure also shows the historical values and forecasts for coal prices in nominal dollars, which is an outlier quantity regarding many aspects of the analysis.

Additional data adjustments. Some data required unit or definition adjustments to be consistent over the entire analyzed time frame. Typically, these adjustments needed to be made on reference case and scenario projections alike.

Constant dollar prices were converted to 2013 dollars for the analysis. In some instances, nominal dollar price projections needed to be converted using constant dollar price projections or vice versa by EIA's inflation rates, which were given in the AEO reports or inferred from prices that were reported both in constant and nominal dollars.

Since we analyze **oil production** and **petroleum consumption** in million barrels per day, some of the projected values had to be inferred from values provided in million barrels per year.

Natural gas prices were initially reported as the average lower 48 wellhead price in dollars per thousand cubic feet (AEO 1982-2012). Later AEOs replaced this with Henry Hub spot prices in dollars per million Btu. We converted million Btu into thousand cubic feet with the heat content for dry natural gas reported in the respective AEO. We did not take data from the most recent Retrospective Review (released in 2015) for natural gas prices, since it lists natural gas prices to electric generating plants instead of wellhead prices.

We work with **coal prices** in constant or nominal dollars per million Btu. While most of the AEOs report coal prices in these units, some of them in addition include projections in a mass-based unit of dollars per short ton. AEOs 1983-1993 report coal prices in dollars per short ton. We use approximate heat contents from the outlooks for conversion. Since heat contents vary marginally, and we are not provided a factor for every single forecasted year, we assign the heat content from the nearest forecasted value, or interpolate if the year is between two years with given heat content. We added the waste coal supplied to **coal production** for the projections of AEO 2007-2016. Waste coal was listed separately for these outlooks, while it was included in coal production before. This is consistent with the Retrospective Reviews, except for a discrepancy for the AEO 2013. We chose to use the values directly from in the AEO in this case.

AEO scenario data. Scenario values were taken from the AEO reports. To compute the envelope scenarios, we found the maximum and minimum of all scenarios in every forecasted year and assigned those to what we called high and low scenarios. These resulting envelope scenarios do not correspond to a single projection of the AEO. The scenarios do not include the early release reference cases, but for AEO 2016 we included the "reference case without Clean Power Plan".

The AEO 2009 has been updated after it was published. We work with this updated reference case to find the forecasting errors. The scenarios however have not been updated. This results in a general mismatch between the scenarios and the reference case for AEO 2009, which is why we left it out of the test set.

2. Error metrics

It is common to refer to the deviation of the forecast from the actual value as *error*. The EIA for example uses this term in its Retrospective Reviews [2]. We work with the relative error for most quantities and transform the relative error for the price quantities into a log-error, which results in a distribution of price forecast errors that is closer to a normal distribution. The analysis is conducted entirely in the relative and log-error metric, but absolute errors could also be used.

Relative errors. We focus on the relative error or percent error in this analysis, because it enables a comparison between forecasts of different quantities. This choice of error however comes with the typical scaling issues of the percentage metric. It is defined as $\epsilon_{rel} = \frac{\hat{y}-y}{y} = \frac{\hat{y}}{y} - 1$, where $\hat{y}$ refers to the forecast and y to the actual value, or observation. The relative errors for all quantities considered in this analysis are displayed in Fig. S4. This is the full set of error samples, also containing the horizons we chose to exclude from the analysis because of their lower sample size. The evolution of the errors over the AEO release years, shown in this figure, makes it easy to identify similarities between the quantities. We can for example see, that electricity price forecast errors look very similar to those of coal price forecasts. In this figure, a large vertical spread indicates that those particular AEO years have resulted in large errors across different horizons. Errors of a similar magnitude over several AEO release years that give the impression to be lined up are in most cases from the same observed value, see for example coal consumption.

We view the forecast densities as a distribution of actual values y around the AEO reference case forecast $\hat{y}$. Also scenarios are treated in this metric. A scenario in our analysis is expressed as the percent error of how much the reference case deviates from that scenario y_S, which is in the resulting relative error metric $\epsilon_{S,rel} = \frac{\hat{y}-y_S}{y_S}$. This means that the errors of high scenarios correspond to $\epsilon_S \leq 0$ and low scenarios to $\epsilon_S \geq 0$. The value of an observation in the relative error metric is computed as $\xi = \frac{\hat{y}-y_{obs}}{y_{obs}}$.

We chose to work with the L_1 loss and mean absolute (percentage) errors instead of the L_2 loss. This means we do

not use the root mean square error (RMSE), which is the risk function (or the expected value) of the L_2 loss. This risk is minimized by the mean. By squaring the errors, L_2 loss inflates the weight of errors that are larger, which is desired if attention needs to be paid to outliers in the data. On the contrary, here we wish to find an estimate of the central point of the distribution that is robust to outliers, which the mean is not. The risk function of the L_1 loss is instead minimized by the median. Especially when faced with a skewed distribution, as it is the case for many of the error distributions in our analysis, the median is a better estimator of central tendency because it is less affected by outliers. In addition, the CRPS reduces to the absolute error (the relative or log error in our case) for a point forecasts, which makes both these metrics compatible.

Log-errors. Prices are typically described as log-normally distributed [4]. In Q-Q-plots of historical AEO price quantities, we found that the logarithm of the prices follows a normal distribution closer than the untransformed prices. This supports the assumption that the prices, even though they are given as an annual average, are approximately log-normally distributed. We make the additional assumption that also the price forecasts follow a log-normal distribution, and introduce an error transformation.

For the transformation, we draw an analogy to logarithmic returns, a concept from financial theory. The return is defined as $r = \frac{\text{future value}-\text{present value}}{\text{present value}}$. If values are log-normally distributed, the log return $\ln(1+r)$ follows a normal distribution[1]. To transform the relative errors for prices, we use very similar arguments where instead of the return we work with the relative error $\epsilon_{rel} = \frac{\hat{y}}{y} - 1$. This results in the log error $\epsilon_{log} = \ln(1+\epsilon_{rel}) = \ln\left(\frac{\hat{y}}{y}\right) = \ln\hat{y} - \ln y$. We compute all of the comparative statistics in ϵ_{log}. We termed the mean absolute log error MALE.

How the loss function changes if the absolute percentage error APE is transformed into the absolute log error, ALE, can be seen in Fig. S1. Here, we define the loss as APE or ALE.

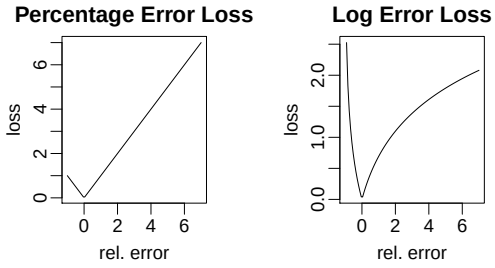


Fig. S1. Comparison of the two types of errors we work with, with APE on the left and ALE on the right. Loss is defined as the absolute error in the respective metric.

3. Summary statistics of the error samples

Normality of the error samples. Here we assess if the errors are normally distributed. Since we use a Gaussian as a parametric density forecast, it is of interest how closely a normal distribution matches the error samples. In addition, we test if the log-errors for the prices are normally distributed, which is the goal of the transformation we apply to price quantities.

We test the assumption that the error samples are normally distributed. We use the Shapiro-Wilk normality test, implemented in the R-package *stats* [5]. The Shapiro-Wilk test is based on the null hypothesis that the sample is normally distributed. The test has the highest power for small sample sizes compared to other tests, even if though the power is fairly low when the true distribution is a symmetric distribution [6]. In Fig. S5, we show the test results for the error samples for two different significance levels, 95% and 99%. We see that for most quantities we cannot reject the null that the errors come from a normal distribution. However, there are some quantities, which with 95% confidence do not have normal errors. In particular, petroleum consumption, coal consumption and total energy consumption exhibit deviations from the normal distribution.

To test the assumption that we should transform the price quantities, we also perform the Shapiro-Wilk normality test on price quantities with transformed errors (Fig. S5). We see that for almost all price quantities, the log-errors are more likely to be normally distributed than untransformed errors. In further analysis not shown here we found that the log transformation has marginal effect on the production and consumption quantities or makes them less Gaussian. Coal price errors are an exception, which for many horizons are bimodal and therefore clearly not Gaussian, even when transformed to log errors. Electricity price errors behave similarly, as electricity prices are correlated to coal prices. How the distribution is adjusted by the log transformation is shown with histograms in Fig. S5. We see that for the example of oil prices, the distribution becomes more Gaussian, whereas the bimodal distribution of coal price errors is largely unaffected by the transformation. Coal prices have been overforecasted for a long period, followed by a period of underforecasting (Fig. S3). This resulted in the bimodal error distribution. We also find that changing the confidence level for rejection of the null hypothesis to 99% allows the error samples of many quantities to appear Gaussian for almost all horizons, with the exception of coal prices.

Autocorrelation. We find that autocorrelation of errors is different from quantity to quantity (Fig. S6). It is typically lower for smaller horizons, larger horizons all show high correlation that only disappears for long lags. Coal prices and electricity prices have a large autocorrelation even for forecasts with small horizons. This matches the pattern that can be observed for coal prices, where we saw long alternating periods of over and underforecasting, and therefore the errors are more correlated (Fig. S3).

In Fig. S4, we can see the autocorrelation reflected in the pattern of errors. This figure, as described above, shows the magnitude of the errors over the release year of the AEO that generated the projection. Where we observe a wave pattern, as for example for coal prices, we find that errors of larger horizons are highly correlated from one AEO to another. This pattern is repeated in electricity prices and transportation energy consumption. Quantities with less autocorrelated horizon samples such as residential energy consumption do not exhibit this pattern. In the case of oil production, we find a relatively

[1] To see this, we start with the definition that if $Z = log(X)$ is normally distributed, X is log normally distributed. So, if $\mathrm{FV} \sim \mathrm{N}$ and $\mathrm{PV} \sim \mathrm{N}$, and noted that $\log(1+r) = \log(1+\frac{\mathrm{FV}-\mathrm{PV}}{\mathrm{PV}}) = \log(\frac{\mathrm{FV}}{\mathrm{PV}}) = \log(\mathrm{FV}) - \log(\mathrm{PV})$ and we know that the sum of normally distributed variables is again a normally distributed variable, we find that $\ln(1+r) \sim N$.

large autocorrelation for small horizons, which can perhaps be attributed to the recent oil and natural gas boom. The observed values changed systematically and rapidly, which was not picked up by many of the recent AEO projections. This is reflected in the waterfall shape of errors for oil production in Fig. S4. Since natural gas production errors were historically larger and more volatile, we do not observe this pattern as clearly here. The pattern of errors that appear lined up, as mentioned in the previous section, does not generally indicate autocorrelation, as this is a result of single outlier observations.

As much as the presence of autocorrelation is a problem for viewing the error series as a random sample, it does not impact the validity of comparing the mean CRPS among the methods. However, for the significance test of improvement of an empirical method over the scenarios, we use the sample of single observation CRPS as a random sample. Here some correlation is to be expected and large correlation could pose a problem. This depends on the autocorrelation of observed values and the AEO forecasts, as well as the similarity of forecast densities from one observation to the other. It is expected to have a similar or lower autocorrelation than the error time series shown in Fig. S6. For our purpose, we assume we can view this autocorrelation as negligible.

Grouping the Quantities. In Fig. S7, we plot the standard deviation of the error samples against the autocorrelation at a lag of $3yrs$ for every horizon separately. This allows us to potentially identify groups of quantities with similar characteristics. The characteristic form seen in the figure does not change much for an autocorrelation coefficient of a different lag. Most apparent is the large variance of errors of the price quantities. We can identify prices with higher autocorrelation (coal and electricity) and with lower (oil and natural gas). This picture emphasizes that the prices form a distinct group among the quantities. In addition, the standard deviation of price errors has a large spread for the different horizons. The electricity price is the most similar to the other quantities outside this group.

The rest of the quantities has a much lower standard deviation, where zooming in on a section of the plot helps to visualize potential differences. We see that the rest of the quantities are fairly similar in these characteristics. Oil production is somewhat different, in that it has a larger standard deviation at a lower autocorrelation coefficient.

From this and the previous analysis we can conclude that treating the price quantities and the other quantities as two distinct groups, and applying the log transformation only to price errors, seems a valuable approach.

4. Details on Density Forecasting Methods

We excluded any historically intractable approaches, i.e. methods, where it is impossible to trace back in retrospect how an analyst would have estimated the uncertainty at the point of decision. A common approach that would fall into that category would be stakeholder elicitation, where the uncertainty range is agreed upon by a number of stakeholders' beliefs about the future. As there is no means of determining how a generic group of stakeholders would have decided at a particular moment in the past, validation and generalization of these types of uncertainty estimates is virtually impossible.

Secondly, we considered but excluded very arbitrary estimates. This could for example be the heuristic of choosing the 10^{th} and 90^{th} percentile as a $\pm 20\%$ error for the forecast five years out. While to our anecdotal knowledge this approach is not uncommon, we chose to exclude it due to the entirely arbitrary nature and the vast number of heuristics that could be employed (e.g., why use 20% and not 15%).

NP_1: Non-parametric density forecasts by retrospective errors. This is a detailed description of the empirical density prediction method NP_1 as introduced in [7]. Methods NP_1, NP_2, and G_1 are based on the assumption that the past forecast errors are a good estimator for the future forecast errors. Under this assumption, the distribution of past errors provides a probabilistic estimate of a future actual value given a point forecast by the same forecaster [8]. For this EPI (NP_1), we use a non-parametric distribution of the errors.

To respect the fact that forecasting gets more and more difficult the further we look into the future, we group the forecast errors by their horizon. For constructing the EPI, we assume that a future forecast error is sampled from the same distribution as past errors. In particular, it is the distribution of all forecast errors with a particular horizon H that determines the uncertainty of the new forecast H years into the future. With the error distributions for a number of consecutive horizons, we obtain a measure for the uncertainty for a time frame $H = 1 \cdots H_{max}$ years into the future. Anchoring the error distribution with $\epsilon = 0$ on the most recent forecast, we obtain a density forecast.

When we create the density forecasts, we need to find the appropriate reconstruction of the predictive density over future real values. In the relative error metric, the statistics of the distribution such as quantiles are reconstructed as actual values y relative to the most recent forecast $\hat{y}$. This is $\epsilon_{rel} = \frac{\hat{y}}{y} - 1 \Leftrightarrow y = \frac{\hat{y}}{\epsilon_{rel}+1}$. When constructing the density forecast from log-errors, we need to use a different expression than if we work with ϵ_{rel}. For log-errors, the density forecast is constructed as

$$\begin{aligned} \epsilon_{log} &= \ln(\hat{y}) - \ln(y) \\ \Leftrightarrow \ \ln(y) &= \ln(\hat{y}) - \epsilon_{log} \\ y &= \exp[\ln(\hat{y}) - \epsilon_{log}] \\ &= \hat{y} e^{-\epsilon_{log}}. \end{aligned} \qquad [1]$$

NP_2: Transforming the errors for the median-centered EPI. For method NP_2, we center the distribution of errors such that the median of the distribution coincides with $\epsilon = 0$. This prevents the density prediction from creating a second point forecast when bias is present in historical forecasts, as it is the case with method NP_1. Here, the goal is to give the largest probability weight to the AEO reference case forecast.

The median-centering is done in percentage points of the errors. This procedure is not based on physical rationale, but it turns out to be a reasonable transformation for small median errors. The centered relative errors are transformed as $\epsilon'_{rel} = \epsilon_{rel} - m_{rel}$. We write ϵ_{ctr} as ϵ' for simplicity. The price forecasts are median-centered in log-errors. Some price quantities have large median errors. If they would be centered in a relative error metric, $\epsilon'_{rel} < -1$ could occur, which is not defined. The log-error metric prevents that situation from occurring.

Centering the error distribution in the log-error metric to ϵ'_{log} changes the relative error as follows below. We center here with the median of the log errors m_{log},

$$\begin{aligned} \epsilon'_{log} &= \epsilon_{log} - m_{log} \\ \ln(1+\epsilon_{rel'}) &= \ln(1+\epsilon_{rel}) - m_{log} \\ 1+\epsilon_{rel'} &= \exp(\ln(1+\epsilon_{rel}) - m_{log}) \\ 1+\epsilon_{rel'} &= (1+\epsilon_{rel})e^{-m_{log}} \\ \epsilon'_{rel} &= (1+\epsilon_{rel})e^{-m_{log}} - 1. \end{aligned} \quad [2]$$

Centering in log-errors retains a crucial property of relative errors, by ensuring that they are defined on the range $-1 < \epsilon_{rel}$. This can be seen by

$$\begin{aligned} \epsilon_{rel'} &= (1+\epsilon_{rel})e^{-m_{log}} - 1 \\ &> (1-1)e^{-m_{log}} - 1 \\ &= -1. \end{aligned} \quad [3]$$

How this change in centering changes the resulting width of the uncertainty interval for a range of errors $-1 < \epsilon_{rel} < 7$ is shown in Fig. S2. We see here as well that centering in the log error space prevents singularities, which can occur when transforming back to the forecast uncertainty when centering in the relative error space.

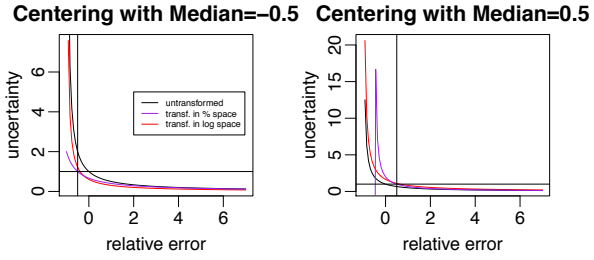


Fig. S2. Comparison of centering in the two error metrics and the impact on calculating the final uncertainty. To the left with a large negative median error and to the right with a large positive median error. We see that the singularity, that occurs when centering in the relative error space, does not occur for centering in the log error space. Median errors are in units of relative and log error respectively.

G_1: List of standard deviations for all quantities. In Table S1, we give the standard deviations of the error samples, which are necessary to implement method G_1 for AEO 2016. We list the standard deviations of the relative errors (or log-errors for prices) for every horizon $H = 0$ to $H = 12$ computed with AEOs 1982-2016.

G_2: Finding the standard deviation of historical values. Method G_2 is a Gaussian uncertainty based on the deviations in the time series of historical values. We find the standard deviation by taking a pair of two historical observations, a horizon H apart, and calculate the relative change of the later value with respect to the earlier value. This is in analogy to the relative error. We find all possible pairs over the time series of historical values for a certain H. The standard deviation of this sample then is the standard deviation that is used to construct the density forecast. For price quantities, we determine the deviation as a log error and then find the standard deviation of those log errors. There is no value for $H = 0$, since $H = 0$ not a real forecast horizon. It corresponds to the error that occurs when in a new AEO the past data has been updated.

Alternative density forecasting methods. The most straightforward integrated approach to obtain a probabilistic forecast is to propagate the uncertainty of both initial conditions and model parameters, most commonly achieved using Monte Carlo simulation. Sensitivity to initial conditions, a feature of many nonlinear systems, is a particular challenge for example for numerical weather prediction. One solution is ensemble weather forecasting, whereby a separate scenario is simulated for each initial condition [9]. These simulation approaches do not consider model misspecification, where the model structure is erroneous, and results depend on the modeler's assumptions about the (future) distribution of input parameters. In the particular case of the AEO projections and the NEMS model, a report by the National Research Council (1992) [10] has recommended the use of multiple probabilistic techniques including Monte Carlo methods and closed-form statistical approaches. They emphasized the need of having reduced-form modules available for shorter run times. The implementation of those methods, however, is considered difficult and might not be feasible. The EIA more recently published a working paper about the use of dynamic stochastic general equilibrium (DSGE) models [11], where the author writes "DSGE models do explicitly incorporate uncertainty and are predominantly forward looking. These models use rational expectations, which imply that consumers are correct on average in forming their expectations about the future values of variables. DSGE models cannot be made very large due to the incorporation of uncertainty, and this limits their usefulness in detailed policy analysis. Their primary uses to date have been in the research work at universities and central banks. Some recent progress has been made in using DSGE models to forecast different macroeconomic variables, but this is an emerging research area."

Other probabilistic forecasting methods are generally very different from the EIA's current forecasting approach, but could in principle give guidance to the AEO scenario selection. Modeling time series data as a stochastic process and methods related to vector autoregressive (VAR) models are common in finance and economics [12]. VAR models might be more suitable for short-term forecasts in the EIA context [11]. There are Bayesian methods that allow for probabilistic forecasting such as Bayesian vector autoregression [13] or Bayesian hierarchical models [14]. In general, many statistical and machine learning methods, such as neural networks [15, 16], can generate density forecasts [17]. When subjective prediction is assessed by expert elicitation, typically the entire predictive distribution is elicited [18]. In principal, an expert elicitation protocol could be modified to quantify the uncertainty around a given point forecast.

5. Sensitivity of the method ranking

Normalizing the CRPS. We normalize the average CRPS for every horizon by the average CRPS for every horizon of the scenario ensemble. This is preferable over normalizing every single observation first, since this would unnecessarily bias the result. To illustrate this, we consider a sample of two instances producing the scores for the alternative density forecast

$CRPS_{Alt} = \{1, 2\}$ and the scenarios $CRPS_S = \{2, 1\}$. We would in this case like to have an average normalized score of $CRPS_{mean,norm} = 1$. By normalizing for every observation, we would obtain $CRPS_{mean,norm} = mean(\{\frac{1}{2}, 2\}) = 1.25$. However, if we normalize the means, we get $CRPS_{mean,norm} = mean(\{1, 2\})/mean(\{2, 1\}) = 1$.

Main ranking method. To find the best density prediction method for each quantity, we rank the average CRPS after normalizing it by the average CRPS_S of the scenario ensemble. We refer to the scenario methods with the subscript S. For every quantity we then average over a core range of horizons $H = 2$ to $H = 9$, and rank these aggregated scores. The method with the lowest average rank is considered the best density over the test range for a given quantity.

We chose to exclude $H = 0$ and $H = 1$ from the core range of horizons because for most forecast users only future values are relevant. The number of observations per horizon in the test range without AEO 2009 ranges from 11 ($H = 0...2$) over 5 ($H = 9$) to 2 ($H = 12$). We exclude the horizons with a sample size smaller than 5 from the core range, which then is $H = 2$ to $H = 9$.

Table S2 summarizes the ranking results for every quantity. It compares the best and second best method of the main ranking procedure, as well has the best method if we average over the larger range $H = 1$ to $H = 12$, employ an alternative ranking method detailed in the next section, or change the test range. We find that the respective best methods do not change much with this sensitivity analysis. Some quantities are however very sensitive to changes in the range of observations since for those quantities two or more methods have very similar scores. For example for natural gas prices and natural gas consumption, the best and second best methods switched after we added the 2015 observation with publication of the AEO 2016. The update of the 2014 observation in AEO 2016 did not have an effect. Those three quantities are an example where the difference is very small. We also see sensitivity for natural gas prices in constant dollars when we remove the first test AEO 2003. The table also lists how much the average normalized CRPS of the best method is lower, and therefore better, than the second method.

Alternative ranking method. To explore the sensitivity of our results for the best density prediction methods for each quantity, we introduce an alternative ranking method. We rank the average CRPS results for each forecasting horizon separately. For every quantity we then average the rank of a method over $H = 0$ to $H = 9$, which results in the final ranking score. The method with the lowest average rank is considered the best density forecasting method for a given quantity. This approach is agnostic about how much the CRPS is improved by a given method over the other. This is the reason why we decided not to use this ranking procedure as the default.

We find that that the method rankings do not change much with the choice of ranking method. The results are listed in Table S2. The alternative ranking method ranks the second best method differently to the main ranking method for only one quantity.

6. Improvement over scenarios

In Fig. S9 we show that we can find a density forecasting method that has a lower mean CRPS than the scenarios for all of the quantities. The only partial exception is petroleum consumption, where that is only true for lower horizons.

Hypothesis test with bootstrap. It is insufficient to know that the aggregated mean CRPS, which we used to rank the methods, is smaller than the aggregated mean CRPS for the ensemble scenarios. Even though a mean might indicate an improvement, the improvement might come for a small fraction of the analyzed observations.

We use a bootstrap method to test how robust, or significant, the indicated improvement is. For each horizon, every observation generates a single CRPS. We resample these scores from the CRPS sample, which depending on the horizon can contain up to 11 elements. We assume complete independence, which means that we do not resample by observation year or make other assumptions about correlation. Under the null hypothesis we assume that the scenarios are the better forecast, i.e. they have the lower aggregated mean CRPS. We test this for every one of the four empirical methods and for every quantity. We resample simultaneously the scores of both the empirical method and the scenario method, which belong to the same observation. We normalize the new mean CRPS by the new mean CRPS_S for every horizon. Averaged over the core horizon range, we obtain a new aggregated normalized mean CRPS. We repeat this a thousand times to find the number of cases where the empirical method could be qualified as worse than then ensemble scenarios, meaning the normalized CRPS is larger than 1. We want this proportion to be smaller than our confidence level of 0.05 to speak of a significant improvement of the empirical method over the ensemble scenarios for the test range.

For all of the quantities, the respective best method is always significantly better than S. Besides performing the hypothesis test for the best methods, we also compared each of the single methods to S. We found that most performed significantly better for all quantities with the exception of NP_2 for constant oil prices which was better with a 92% confidence, and NP_1, which only performed significantly better for six of the eighteen quantities.

We also compared the best methods with the SP_1 method (Gaussian based on scenarios), and found that we can be 95% confident for almost all of the quantities that we found a significantly better uncertainty estimation method for the test range. The only exception is petroleum consumption, where the best method is only better at 74% confidence.

Further analysis of the scenarios. To understand if the scenario range is too narrow, we measure the coverage probability of the range between the envelope scenarios. This corresponds to the percentage of observations that were lower than the highest and higher than the lowest scenario for test AEO 2003-2014, without AEO 2009 (Fig. S10). We find that the coverage varies for different quantities and for different horizons, but it is generally very low with an average of 13.7%. This average is for the core horizon range and all quantities. Typical prediction intervals are intended to cover for example one or two standard deviations of a Gaussian distribution, which correspond to about 68% and 95% respectively.

We note that EIA's AEO scenarios are not intended to have a certain coverage probability. They are sensitivity cases on certain input assumptions. Since only one or very few assumptions, such as the impact of a particular policy, are

changed at a time, the side cases typically do not differ as much from the reference case as they would if several assumptions were changed simultaneously. If the scenario range would be used for communicating the uncertainty, several assumptions would need to be changed simultaneously and probabilities would need to be attributed. Nevertheless, the EIA writes for example in its most recent AEO 2017 [1] "EIA addresses the uncertainty inherent in energy projections by developing side cases with different assumptions of macroeconomic growth, world oil prices, technological progress, and energy policies." In our analysis, we use the SP_1 method to account for a wider uncertainty based on the scenarios. The method uses the range to the widest envelope scenario (of both low and high) as one standard deviation to fit a Gaussian distribution with the reference case as the mean. In this case, the observation is expected to be within that range only 68% of the times, which is a lenient interpretation of the scenarios, particularly considering that the scenario range is often asymmetric.

7. Point forecast comparison

We compare the mean absolute percentage/log error (MAPE/MALE) of three alternative point forecasts with the AEO reference case, similar to the CRPS significance test. Point forecast comparison allows us to understand that even though in some cases it is better to correct the best estimate forecast with the bias of the EPI, in most cases the AEO and therefore a centered error distribution performs better over the test range AEO 2003-2014 without AEO 2009. We exclude AEO 2009 to make the results consistent with the density forecast. In Fig. S8, we show the results for all quantities. From the fact that the reference case seemed to be the better forecast than the median of NP_1 for all quantities but two, we could anticipate that NP_1 would not create a good empirical density forecasts. This was a reason to introduce the centering technique of method NP_2.

Point forecast results. The point forecast comparison is designed to compare the median of the error distribution (bias) to the AEO reference case. In addition, we compare the reference case to two benchmark forecasts. Persistence is the last observation, or here the $H = 0$ forecast. Over the test range it was better than the reference case for 10 of the 18 quantities. This surprisingly good result is probably particular to the recent historical evolution of many quantities. It remains to be analyzed over other test ranges. Another point forecasting method is an interpolation of a simple linear regression over a fixed window. The length of the window has been optimized for the test range, excluding AEO 2009. We tested a window of 5 to 10 years and found that a window of 7 years shows a better forecast for the largest number of quantities, which is 8. This is based on an optimization both on the data pre AEO 2016 and the data updated with AEO 2016. The optimal window range does not change if AEO 2009 is included, but the simple linear regression generally performs worse.

Significance of Point Forecast Performance. We use a similar hypothesis test with bootstrap for the point forecast performance, as described in the previous section for the density forecast performance. Instead of the normalized CRPS, here we normalize the MAPE/MALE as $MAPE_{norm} = \frac{MAPE_{method}}{MAPE_{AEO}}$. We test if this quantity is significantly below 1, which would mean that the alternative method performed better over the test range than the AEO reference case. As before, we resample the absolute percentage error or log error samples for every horizon, and then average to get a MAPE/MALE for every horizon. We then normalize this average and determine the mean over the core horizon range $H = 2$ to $H = 9$. If less than 5% of the values are > 1, we speak of a significant improvement of the method over the AEO reference case for that particular quantity.

8. Analysis omitted in the final paper

To evaluate the calibration of the density forecasts, we also produced probability integral transform (PIT) values [19]. The PIT is defined as the value of the predictive CDF that an observation would have. A fundamental property of this variable is that it has a standard uniform distribution, if the historical value is sampled from a distribution that is equal to the density forecast. To assess if the density forecast is well-calibrated over all forecasts and all horizons we can determine if the PIT are sampled from a standard uniform distribution and if they are independent and identically distributed (iid) [19]. We used the Kolmogorov-Smirnov test to compare the distribution of PIT with the standard uniform distribution, and assessed the autocorrelation of the PIT time series. While this procedure was a good visual tool to understand the calibration of the density prediction, it was however not an adequate option to compare different methods quantitatively. We therefore discarded this method in favor of the CRPS.

We have also tried uncertainty estimation methods that weigh the errors depending on their expected relevance for future errors, considering the non-stationary nature of the error time series. We considered a nearest neighbor weighting method and a method that identifies intervals between non-stationarities and assigns weights accordingly. Those methods, however, have only in some cases improved method NP_1 and did not perform as expected. We believe that this approach could be more promising for forecasting problems with more data.

1. U.S. Energy Information Administration (2016) Annual Energy Outlook. http://www.eia.gov/forecasts/aeo/.
2. U.S. Energy Information Administration (2015) Annual Energy Outlook Retrospective Review. https://www.eia.gov/forecasts/aeo/retrospective/.
3. Anderson AT (1996) Differences between Energy Information Administration Energy Forecasts: Reasons and Resolution. *Energy Information Administration: Iusses in Midterm Analysis and Forecasting*.
4. Sprenkle CM (1961) Warrant prices as indicators of expectations and preferences. *Yale Economics Essays* 1:178–231.
5. R Core Team (2015) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).
6. Razali NM, Wah YB (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics* 2(1):21–33.
7. Williams WH, Goodman ML (1971) A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association* 66(336):752–754.
8. Lee YS, Scholtes S (2014) Empirical prediction intervals revisited. *International Journal of Forecasting* 30(2):217–234.
9. Taylor JW, McSharry PE, Buizza R (2009) Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion* 24(3):775–782.
10. Council NR (1992) *The National Energy Modeling System*. (The National Academies Press, Washington, DC).
11. Arora V (2013) An evaluation of macroeconomic models for use at eia.
12. Tay AS, Wallis KF (2000) Density forecasting: a survey. *Journal of forecasting* 19(4):235–254.
13. Cogley T, Morozov S, Sargent TJ (2005) Bayesian fan charts for uk inflation: Forecasting and sources of uncertainty in an evolving monetary system. *Journal of Economic Dynamics and Control* 29(11):1893–1925.
14. Raftery AE, Li N, Ševčíková H, Gerland P, Heilig GK (2012) Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences* 109(35):13915–13921.

15. Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software* 15(1):101–124.
16. Hippert HS, Pedreira CE, Souza RC (2001) Neural networks for short-term load forecasting: A review and evaluation. *Power Systems, IEEE Transactions on* 16(1):44–55.
17. Gneiting T, Katzfuss M (2014) Probabilistic forecasting. *Annual Review of Statistics and Its Application* 1:125–151.
18. Diebold FX, Tay AS, Wallis KF (1997) Evaluating density forecasts of inflation: The survey of professional forecasters in *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, eds. Engle RF, White H. (Oxford University Press).
19. Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts, with applications to financial risk management. *International Economic Review* 39:863–883.

Table S1. Standard deviations of the forecast errors from AEO 1982-2016

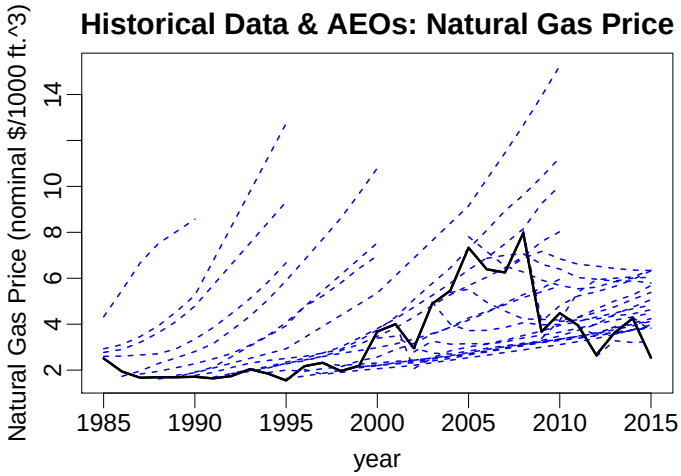
Quantity	H=0	H=1	H=2	H=3	H=4	H=5	H=6	H=7	H=8	H=9	H=10	H=11	H=12
Oil Price	0.029	0.227	0.357	0.423	0.530	0.577	0.668	0.754	0.823	0.893	0.965	1.003	0.988
Oil Price (const.)	0.060	0.229	0.352	0.408	0.505	0.545	0.628	0.703	0.754	0.808	0.865	0.893	0.874
Petroleum Cons.	0.009	0.022	0.038	0.053	0.060	0.070	0.079	0.092	0.105	0.120	0.131	0.139	0.144
Oil Production	0.016	0.050	0.086	0.113	0.131	0.136	0.135	0.133	0.125	0.132	0.144	0.160	0.167
Natural Gas Price	0.051	0.219	0.353	0.428	0.534	0.608	0.673	0.717	0.762	0.761	0.794	0.804	0.770
Natural Gas Price (const.)	0.065	0.218	0.348	0.418	0.518	0.586	0.645	0.679	0.712	0.705	0.725	0.730	0.701
Natural Gas Cons.	0.021	0.042	0.062	0.075	0.083	0.096	0.107	0.114	0.116	0.124	0.129	0.127	0.129
Natural Gas Prod.	0.019	0.040	0.061	0.075	0.090	0.104	0.116	0.124	0.127	0.132	0.129	0.126	0.116
Coal Price	0.060	0.076	0.133	0.187	0.246	0.303	0.362	0.421	0.481	0.535	0.585	0.624	0.641
Coal Price (const.)	0.037	0.076	0.125	0.169	0.220	0.268	0.317	0.365	0.410	0.451	0.486	0.514	0.525
Coal Consumption	0.020	0.045	0.062	0.078	0.097	0.123	0.146	0.162	0.174	0.188	0.190	0.197	0.207
Coal Production	0.019	0.039	0.054	0.059	0.068	0.081	0.092	0.102	0.107	0.117	0.116	0.121	0.130
Electricity Price	0.026	0.049	0.085	0.112	0.142	0.167	0.190	0.214	0.240	0.262	0.285	0.304	0.315
Electricity Sales	0.008	0.015	0.023	0.031	0.037	0.044	0.051	0.059	0.068	0.076	0.080	0.086	0.090
Total Energy Cons.	0.008	0.019	0.028	0.034	0.041	0.051	0.060	0.069	0.080	0.091	0.098	0.103	0.108
Residential Energy Cons.	0.025	0.042	0.039	0.038	0.040	0.048	0.056	0.057	0.064	0.073	0.074	0.076	0.078
Commercial Energy Cons.	0.021	0.033	0.042	0.052	0.056	0.059	0.069	0.078	0.087	0.100	0.103	0.109	0.108
Transportation	0.017	0.026	0.038	0.050	0.065	0.080	0.095	0.111	0.127	0.134	0.150	0.162	0.169

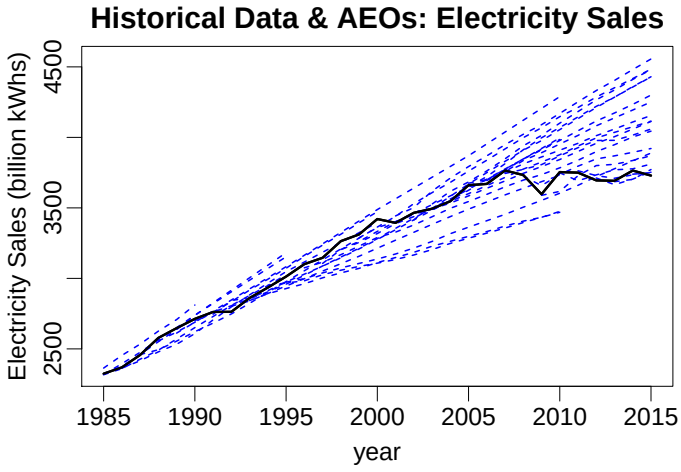
SD are given as ϵ_{rel} except for the price quantities, which are given as ϵ_{log}. These can be used to construct a Gaussian density with quantile y around a forecast $\hat{y}$, which is defined as $y = \frac{\hat{y}}{\epsilon_{rel}+1}$ or $y = \hat{y}e^{-\epsilon_{log}}$ for relative errors and log errors respectively. Values are subject to change as historical values are updated or additional AEOs are released.

Table S2. Ranking results and sensitivity analysis for every quantity

Quantity	best	second best	$\frac{2^{nd}b.-best}{best}$	with AEO 2009	$H = 1$ to 12	test AEO 2004-2014	no obs. 2015	alt. ranking
Oil Price (nominal $)	G2	G1	0.8%	G2	G2	G2	G2	G2
Oil Price (constant $)	G2	G1	2.3 %	G2	G2	G2	G2	G2
Petroleum Cons.	G2	G1	1.8 %	G2	G2	G2	G2	G2
Oil Production	G1	NP2	4.1 %	G1	G1	G1	G1	G1
Natural Gas Price (nom. $)	G1	G2	0.8 %	G1	G1	G1	G2	G1
Natural Gas Price (const. $)	G1	G2	1.0%	G1	G1	NP1	G2	G1
Natural Gas Consumption	G1	G2	0.2 %	G1	G1	G1	G2	G1
Natural Gas Production	G1	NP1	2.9 %	G1	G1	G1	G1	G1
Coal Price (nom. $)	NP2	G1	6.5%	NP2	NP2	NP2	NP2	NP2
Coal Price (const. $)	NP2	G2	9.0%	NP2	NP2	NP2	NP2	NP2
Coal Consumption	G1	NP2	0.9%	G1	G1	G1	G1	G1
Coal Production	NP1	G2	12.7%	NP1	NP1	NP1	NP1	NP1
Electricity Price	NP2	G1	6.5%	NP2	NP2	NP2	NP2	NP2
Electricity Sales	G1	G2	2.1%	G1	G1	G1	G1	G1
Total Energy Cons.	NP2	G2	1.9%	NP2	NP2	NP2	NP2	NP2
Residential Energy Cons.	NP1	G2	7.2 %	NP1	NP1	NP1	NP1	NP1
Commercial Energy Cons.	G1	NP2	1.3%	G1	G1	G1	G1	G1
Transportation	G1	NP2	4.7 %	G1	G1	G1	G1	G1

Ranking results and sensitivity. The improvement of the best forecasting method with respect to the second best is measured in percentage difference of the normalized average CRPS. The best methods from various sensitivity analyses are listed to the right. We vary one assumption at a time. Deviations from the default ranking results are indicated in blue. The default ranking is performed on AEOs 2003-2014 without AEO 2009, observations 2002-2015, and over horizons $H = 2$ to $H = 9$.





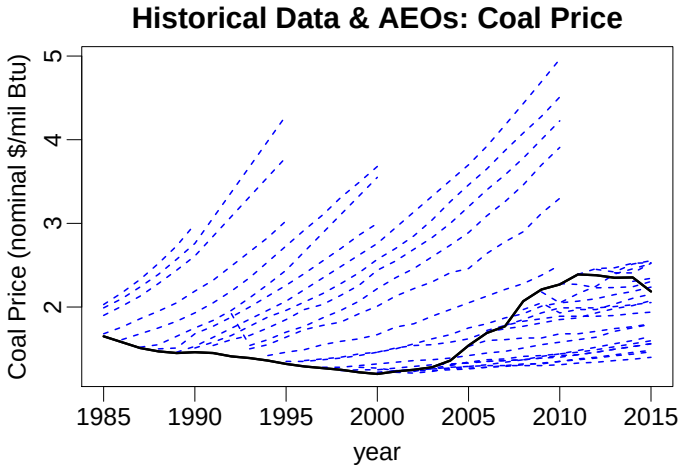


Fig. S3. The historical values and AEO projections for the example quantities natural gas wellhead prices and total electricity sales, and the outlier case coal prices to electric generating plants. The black solid line indicates the historical yearly averages as listed in the EIA Retrospective Reviews. The annual projections from the AEOs 1982-2016 are shown in blue dashed lines. The unusual coal price projection for 1992 in AEO 1993 is not an error in the data.

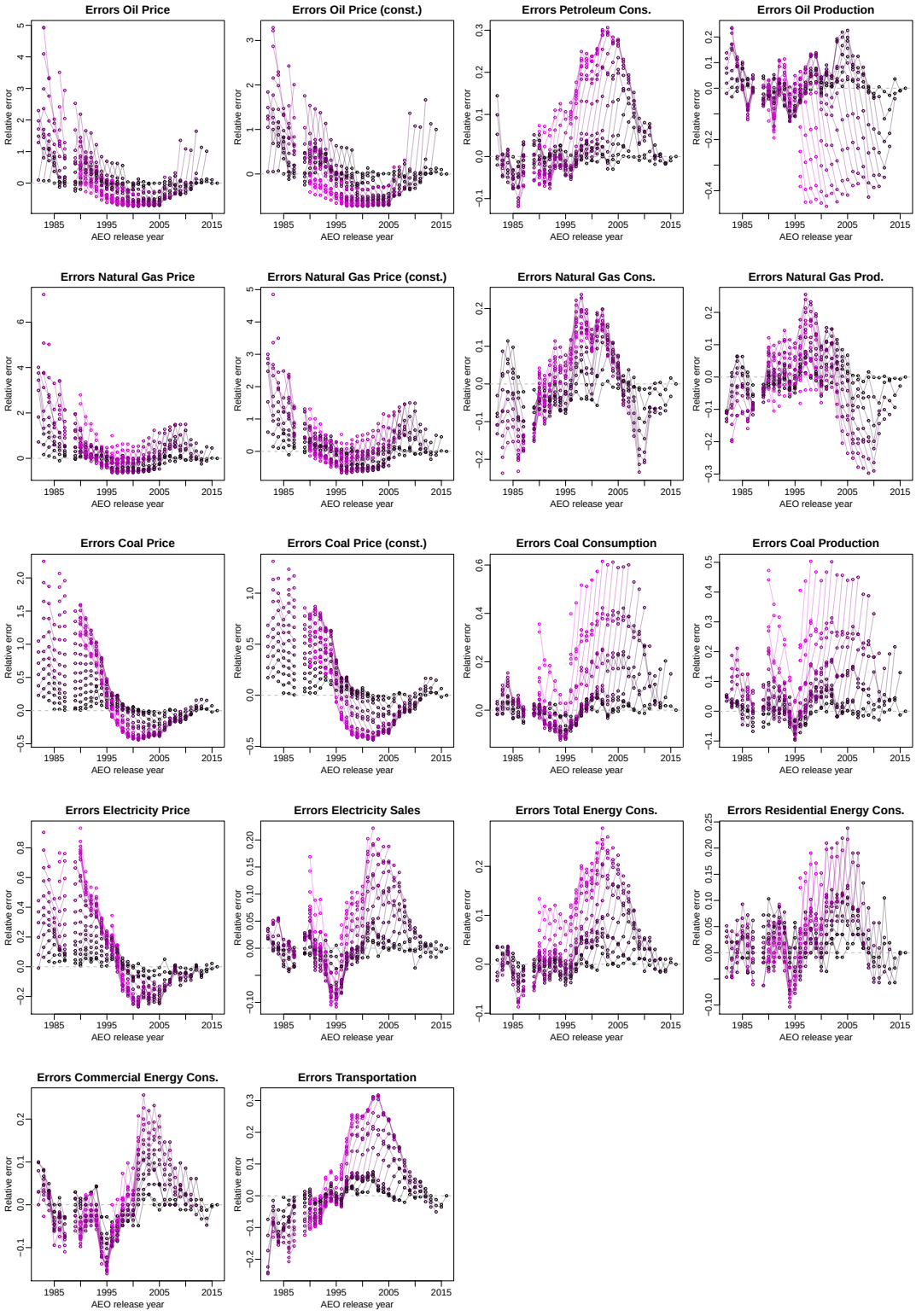


Fig. S4. The relative errors in this data set for all quantities. Each color connected with a line corresponds to a horizon, ranging from $H = 0$ in black to $H = 21$ in purple. The price forecast errors are untransformed.

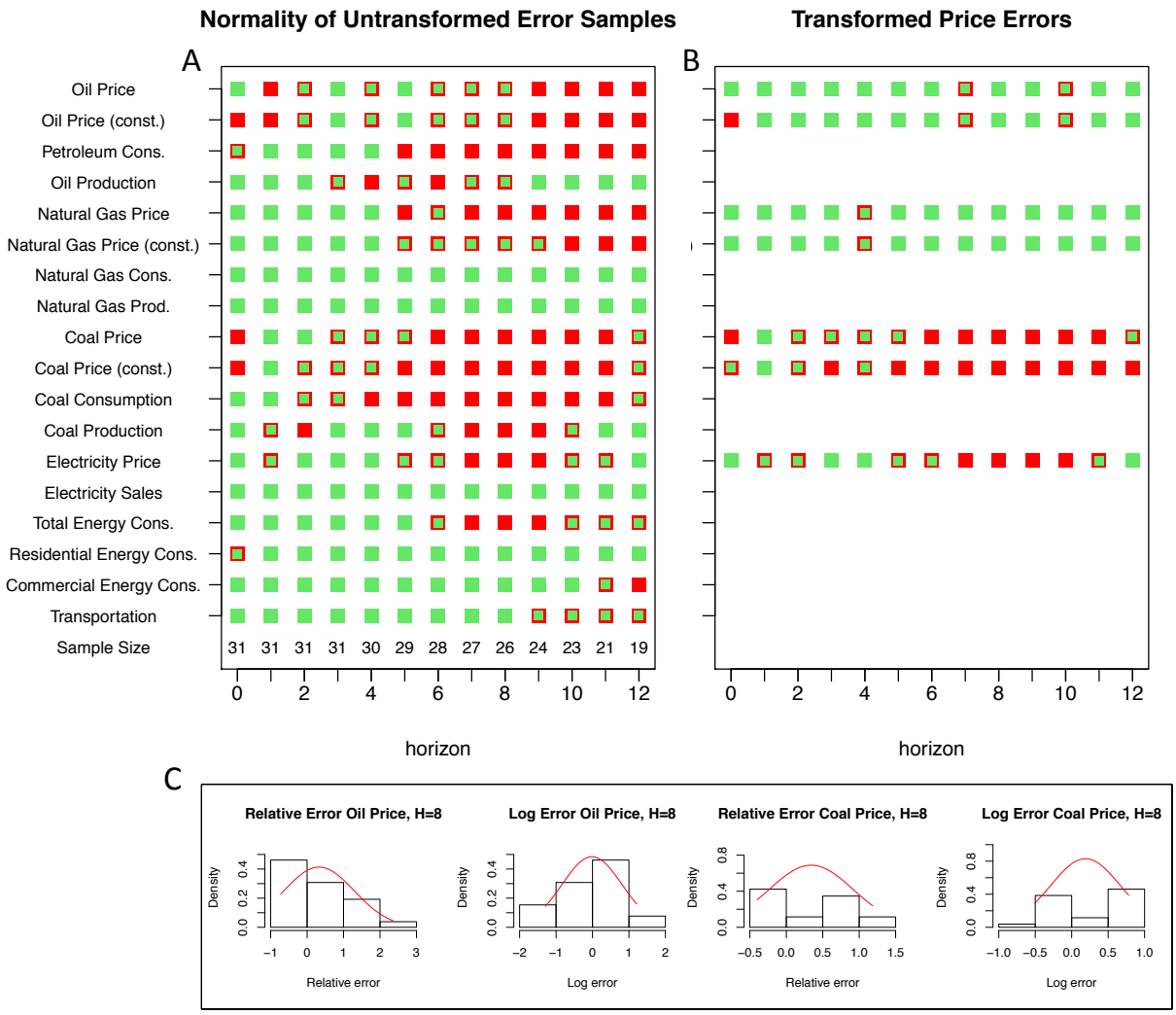


Fig. S5. The results of the Shapiro-Wilk Normality Test with the original relative errors (A) and the transformed errors for the price quantities (B). Red indicates that the sample is not normally distributed with a certain confidence, while green corresponds to those samples where the null hypothesis of a normal distribution cannot be rejected. The underlying larger square corresponds to rejection with confidence $\alpha = 0.05$, and the smaller to $\alpha = 0.01$. (C) Two example histograms of untransformed and transformed errors with Gaussian fit, illustrating how the log error is much more normally distributed than the relative error for oil prices. The transformation has instead little effect on the bimodal coal prices.

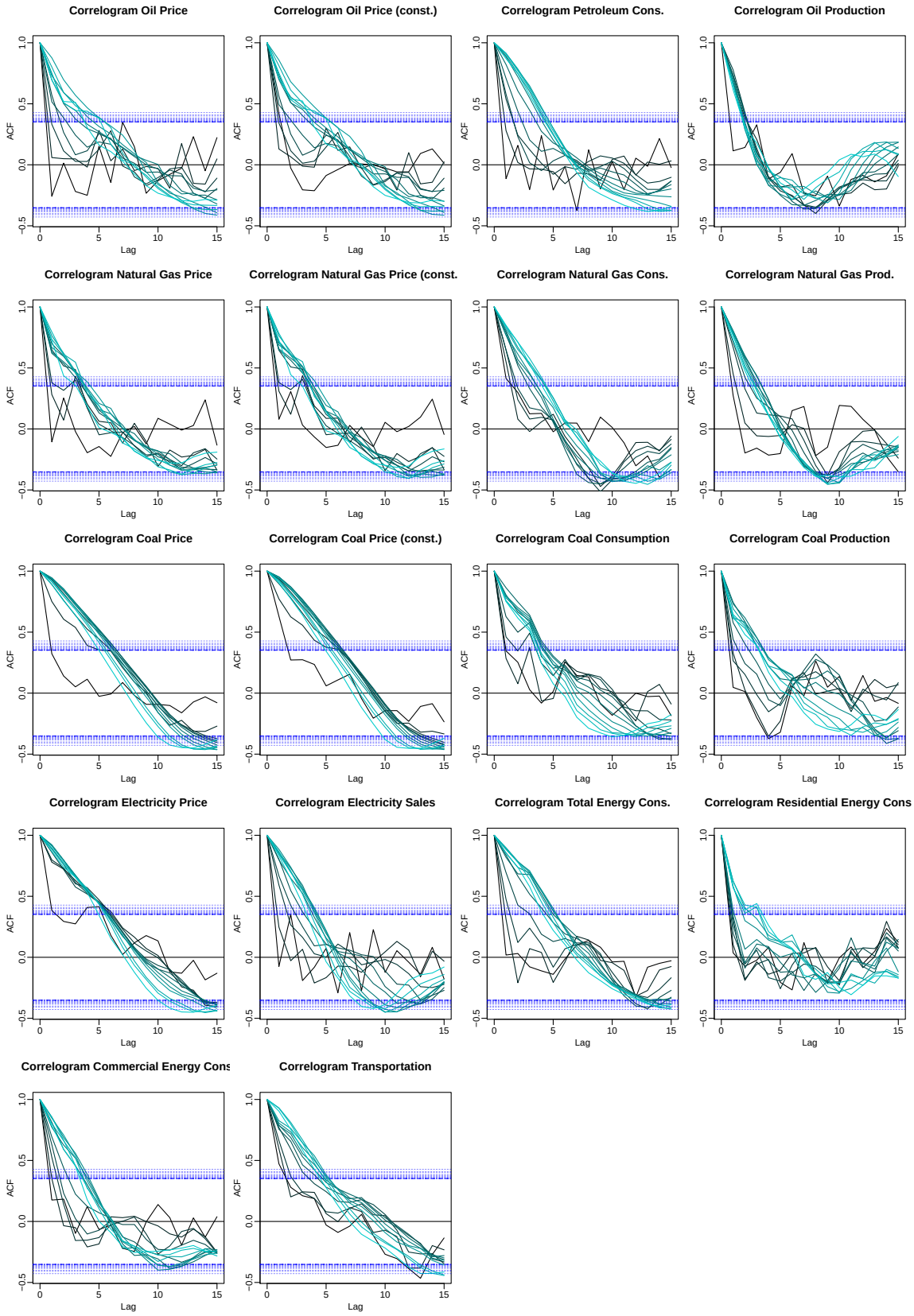


Fig. S6. The correlograms indicating the autocorrelation in the time series of error samples. Every line shows how the error for a given horizon H is correlated to the error for the same H from a previous AEO. Results for different horizons are summarized in the same plot for every quantity. The colors range from $H = 0$ in black to $H = 12$ in light turquoise. The $\alpha = 0.05$ confidence bands for autocorrelation are indicated in dashed blue lines, they vary for different samples sizes. The confidence region is larger for larger H.

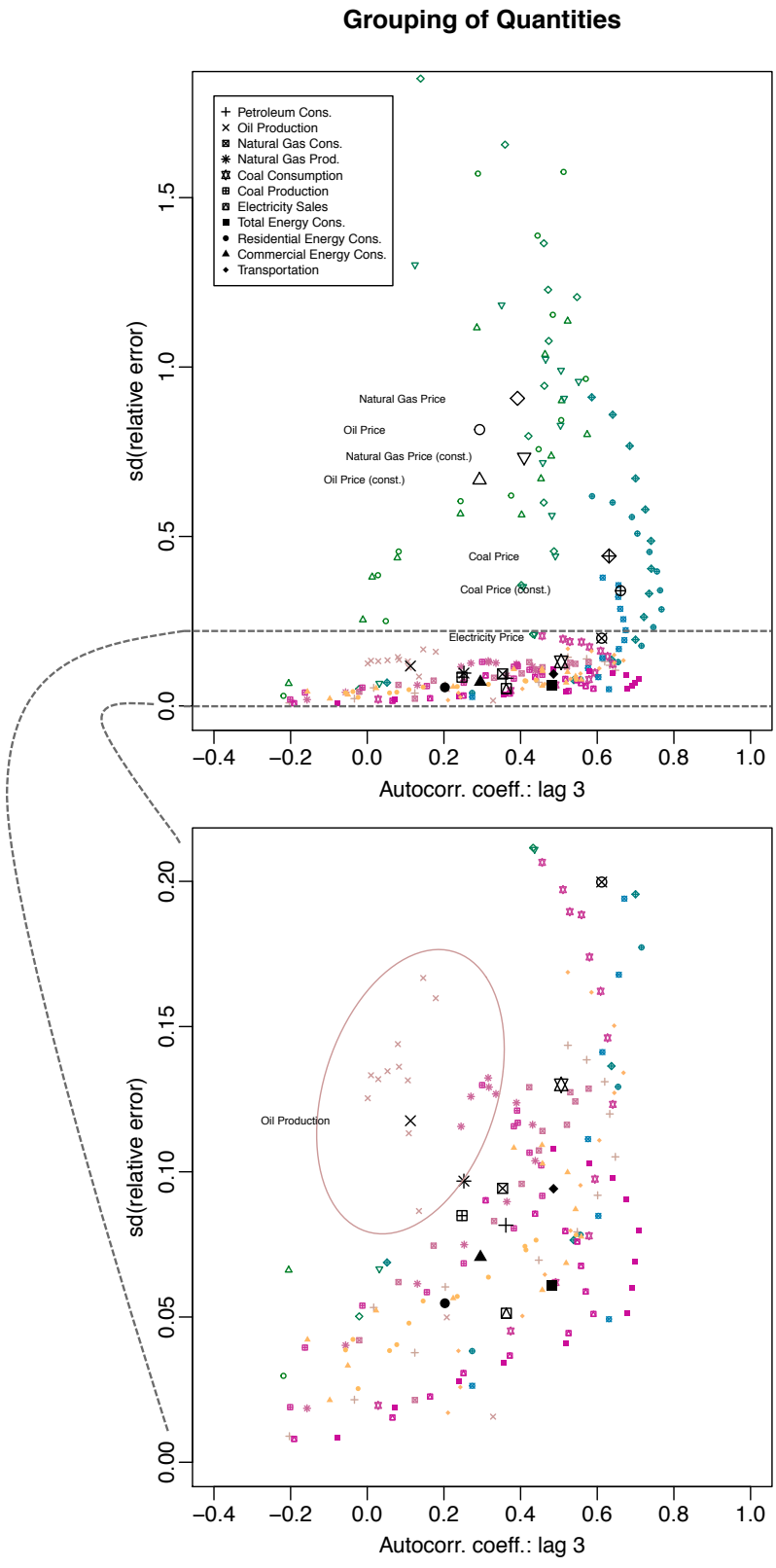


Fig. S7. The standard deviation of relative error samples for distinct quantities and horizons $H = 0 \ldots 12$ plotted against the autocorrelation coefficient at lag 3 yrs. The top image is at full scale while the bottom image is cropped. The colors correspond to the three variable classes of prices (blue/green), production and consumption (magenta), and energy consumption by sector and total (orange). Every color and shape is assigned to one quantity. The black points indicate a mean forecasting error over all horizons for each quantity. Prices form a distinct group in this graph, with a much larger standard deviation than the other quantities. Coal and electricity price errors have a higher autocorrelation and a lower standard deviation than the other price quantities. The ellipsoid in the lower image highlights that oil production distinct from the other quantities, which is due to its low autocorrelation at lag 3 yrs.

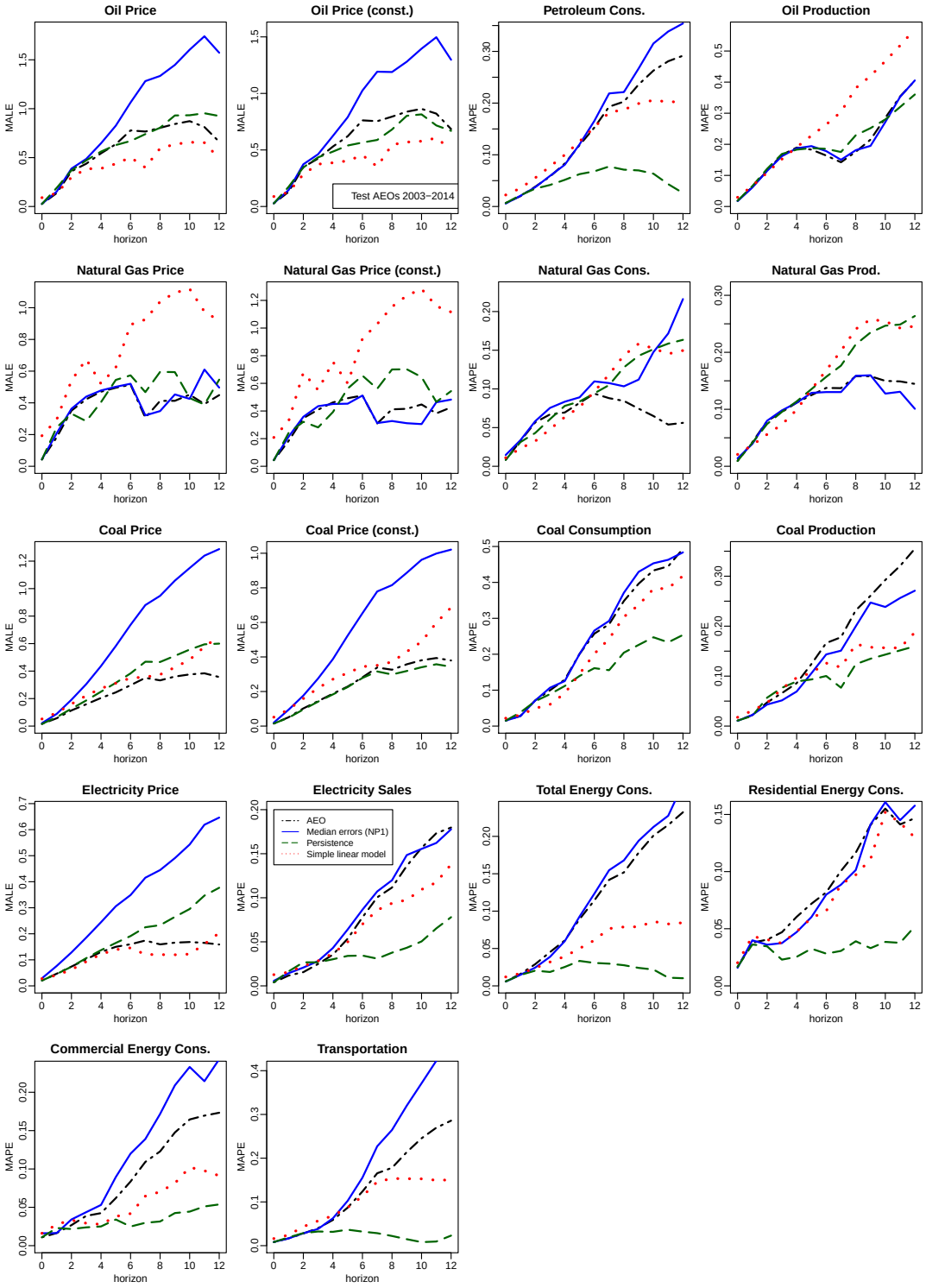


Fig. S8. The results for the MAPE and MALE for all quantities. This is with the test range AEO 2003-2014, and excluding AEO 2009.

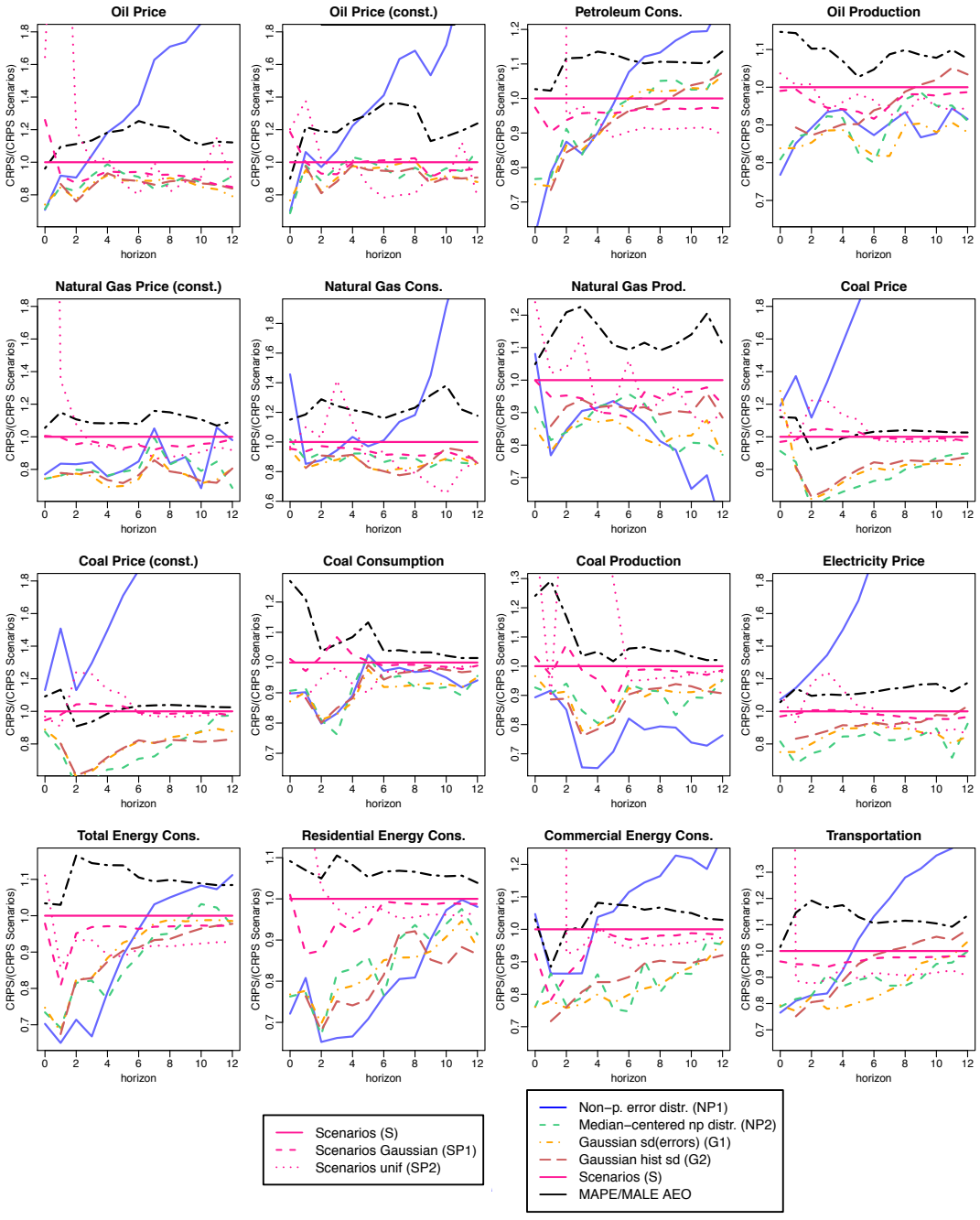


Fig. S9. Relative improvement of the methods with respect to the highest and lowest scenarios for the test range AEO 2003-2014. Values are plotted as fraction of the CPRS of the scenario ensemble (S). A value lower than 1.0 corresponds to a better density forecast. SP_1 corresponds to a normal distribution with the scenario range as 1 SD, and SP_2 is a normalized CRPS of a uniform PDF between the envelope scenarios.

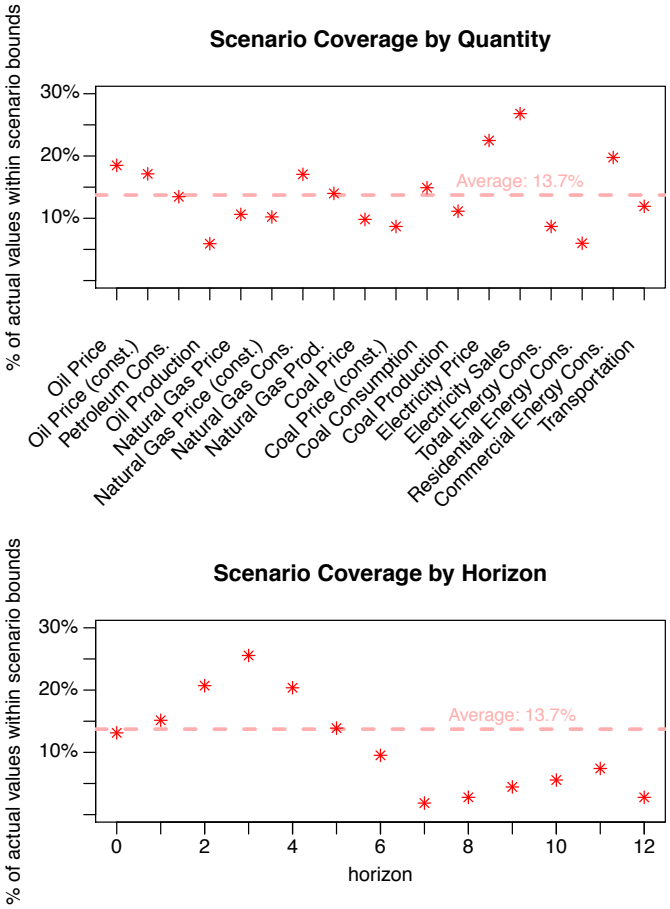


Fig. S10. The coverage probability of the scenario range over the test range AEO 2003-2014 without AEO 2009. The coverage probability refers to the percentage of observed values within the range between the envelope scenarios. The average is computed as the average over $H = 2$ to $H = 9$ for every quantity (shown in A) and then averaged over the 18 quantities. The coverage for every horizon averaged over all 18 quantities is shown in (B).

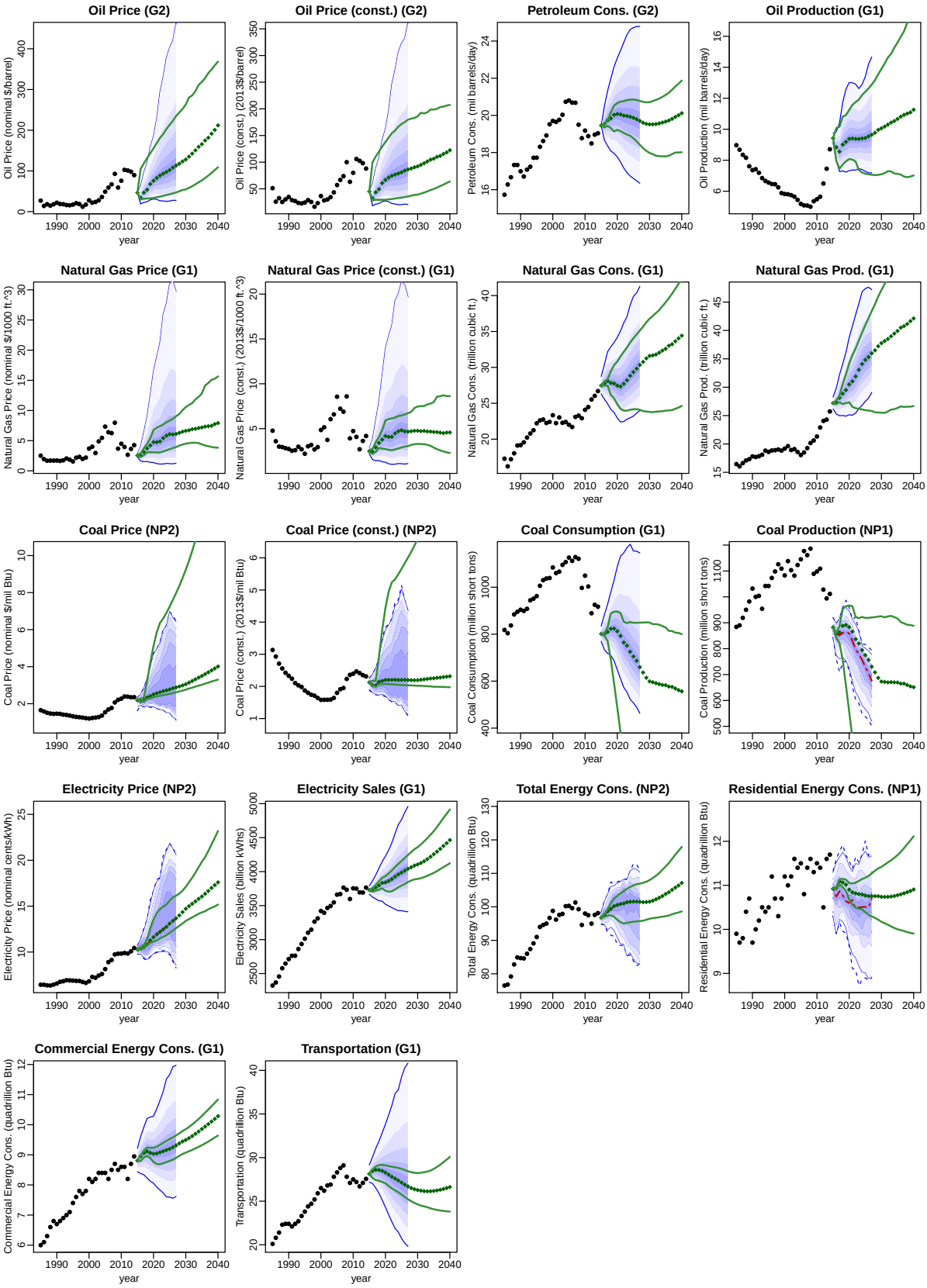


Fig. S11. Density forecasts with the best method for every quantity based on AEO 2016. The different shades correspond to the percentiles 2, 10, 20, 30, ..., 80, 90, 98. The prediction interval can be very large, since it estimates that only 4% for a future value will fall outside of this interval. The red dashed line indicates the median if different from the reference case. The scenario range (in green) changes greatly from one AEO to another and is somewhat correlated to the number of scenarios published, which is why some AEO scenario ranges might be as wide as the empirical uncertainties. AEO 2016 has a large number of scenarios compared to other AEOs.