Carnegie Mellon Electricity Industry Center Working Paper CEIC-16-04 www.cmu.edu/electricity

Empirical Prediction Intervals Improve Energy Forecasting

Lynn H. Kaack[1], Jay Apt[1], M. Granger Morgan[1], Patrick McSharry[2,3]

November 15, 2016

[1]Department of Engineering and Public Policy, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
[2]Smith School of Enterprise & the Environment, Oxford University, South Parks Road, Oxford OX1 3QY, UK
[3]ICT Center of Excellence, Carnegie Mellon University, Kigali, Rwanda

Abstract

Energy projections, such as those contained in the U.S. Energy Information Administration (EIA)'s Annual Energy Outlook (AEO), are important 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, 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 the past forecasting errors, gives good 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 non-parametric empirical density forecasting method. Our findings give guidance on how to evaluate and communicate uncertainty in future energy outlooks and forecasts in other fields.

Introduction

Forecasts of quantities such as electricity and fuel demands, commodity prices, and specific energy consumption and production rates are often used to inform private and public investment decisions and long-term strategies [1, 2, 3]. 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 forecasts are those of the U.S. Energy Information Administration (EIA) and the International Energy Agency (IEA). These forecasts are 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 [4, 5, 6]. Providing information on the likely uncertainty associated with such forecasts would help individuals and organizations use them in a more informed and realistic manner.

All of the energy outlooks mentioned above provide point forecasts without a probabilistic treatment of uncertainty. Often, point forecasts are labeled as a "reference scenario", and are accompanied by alternative scenarios[1]. While scenarios may be used to bound a range of possible outcomes, they can easily be misinterpreted [8] 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 [9, 8, 10, 11, 12]. Shlyakhter et al. criticize the EIA for not treating uncertainty in the Annual Energy Outlook (AEO) [9]. Density forecasting is increasingly becoming the standard [11, 13] in a variety of disciplines ranging from forecasts of inflation rates [14, 15, 16], financial risk management and trading operations [17, 18], to demographics [19], peak electricity demand [20] and wind power generation [21, 22]. There are a number of procedures for probabilistic forecasting [17]. 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 for attaching an uncertainty distribution to a pre-existing point forecast.

The importance of density forecast evaluation has been discussed by several authors [23, 24, 12, 25]. When methods are chosen to generate probabilistic energy forecasts, such evaluation is often omitted. Our work is a step towards making energy density forecasting more feasible and robust by framing how to evaluate a probabilistic forecast in this setting.

[1]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 [7]. 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.

Choosing a density forecasting method

The goal of our analysis is to choose a method that estimates most accurately the uncertainty that should be associated with a future forecast. 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 distribution from which this future observation is drawn.

Density forecasts are evaluated by their calibration and their sharpness subject to calibration [24]. By sharpness we mean that narrower PDFs are preferable. Calibration, as a core concept of 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) [25, 26, 27] 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 non-parametric estimates of forecast uncertainty in the form of probability density functions (PDFs) (see Table 1 and *Materials and Methods*). A simple method of empirical prediction intervals (EPI), first published by Williams and Goodman [28], 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 employed in a number of fields such as meteorology [29], including the creation of the classic "cone of uncertainty" now routinely produced for likely hurricane tracks [30], future commodity prices [31], and the values of macroeconomic variables such as inflation [15]. There is a continuing interest in the method from researchers in

applied mathematics and statistics [13, 32, 33]. We introduce a second non-parametric 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 [34], which are chosen based on EIA's Retrospective Review [35] (*Materials and Methods*). The AEO forecasting record spans more than thirty 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 [33], 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 [4, 36, 1, 3, 37] and the projections by the IEA [5]. Generally, authors have focused on a mean percent error and directional consistency of errors, also termed bias. Shlyakhter et al. [9] 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 forecast. The prediction intervals are also compared to 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

Table 1: Empirical density forecasting methods compared

Method	Parametric	Based on	Median centered in $\epsilon = 0$
NP1	no	forecast errors	no
NP2	no	forecast errors	yes
G1	yes	forecast errors	yes
G2	yes	historical deviations	yes

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 non-parametric method as in [28] (NP_1), performed best in very few cases. The centered non-parametric 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 two of the quantities, natural gas wellhead price in nominal dollars per 1000 cubic ft. (hereafter natural gas price) and total electricity sales in billion kWhrs (hereafter electricity sales) for illustration purposes. Results for all 18 quantities can be found in the *SI*.

Error metric and transformation for price quantities

All forecast evaluation scores are computed on the basis of the deviations of the forecasts $\hat{y}$ with historical values y, referred to as error. We found it useful to work with the percent error, or relative error, $\epsilon_{rel} = \frac{\hat{y}-y}{y} = \frac{\hat{y}}{y} - 1$. Percent 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 [38], we modify the error for the 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*).

The structure of the relative errors as a function of forecast year and forecast

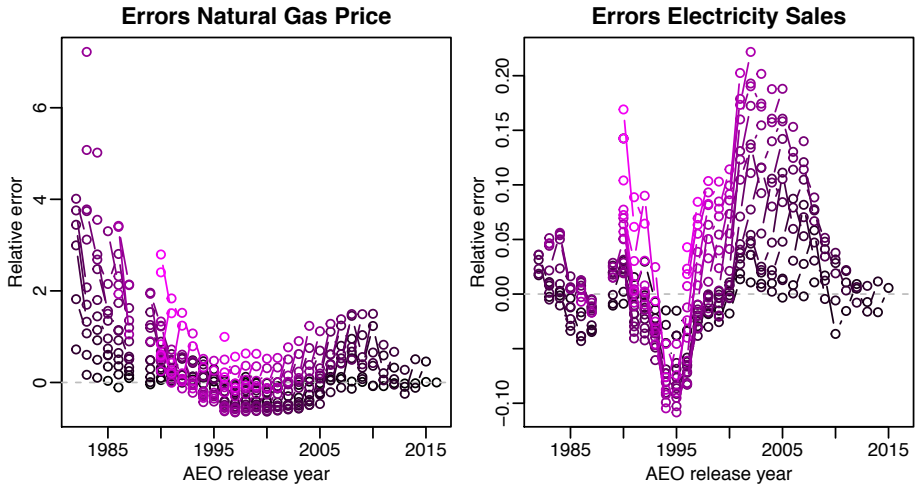

Figure 1: Forecast errors by AEO release year. Differently colored lines correspond to forecast horizons ranging from $H = 0$ in black to $H = 21$ in purple. All forecast errors are untransformed. Note the different scale. No AEO was released for 1988.

horizon is shown in Fig. 1. The horizon H refers to the number of time steps, or years, into the future that the forecast is made. The AEO projections reflect uncertainty in past values, which is why e.g. for AEO 2016 we refer to 2015 as $H = 0$, to 2016 as $H = 1$, and so on.

Retrospective analysis can inform density forecasts

We illustrate examples of the four probabilistic forecasting methods listed in Table 1. Fig. 2 and Fig. 3 compare the non-parametric methods to the methods that performed better for the two example quantities, that is, the two Gaussian predictions.

A non-parametric distribution of the errors (NP_1) results in the EPI shown in Fig. 2. We see that the median of the errors is not exactly zero, which is often referred to as bias. 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 non-parametric 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. 3. In the percent error space, centering is done by subtracting the median error

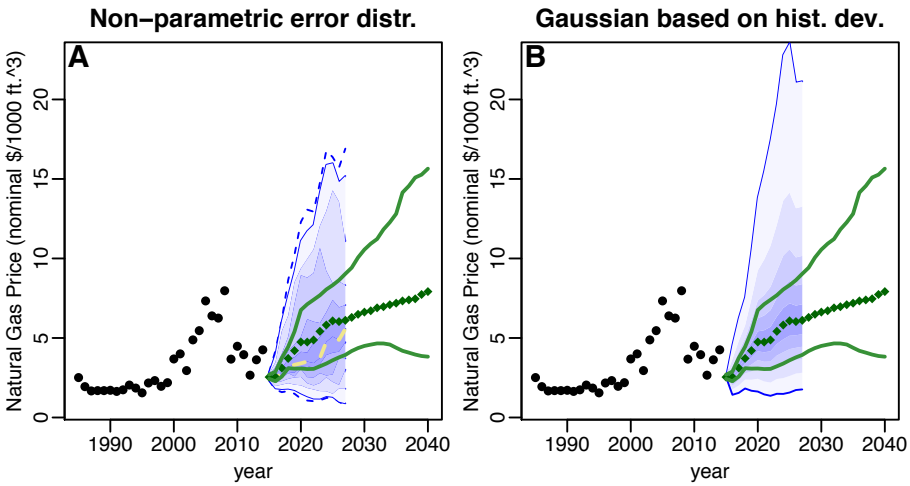

Figure 2: Density forecasts for natural gas prices in nominal $. (A) Non-parametric EPI based on forecast errors (NP_1). (B) Gaussian density forecast based on the variability of historical values (G_2), which tested to be the better estimate. Historical values are indicated by black dots, the AEO 2016 reference case by green diamonds and the density forecast in 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 (in dashed khaki) does not coincide with the reference case.

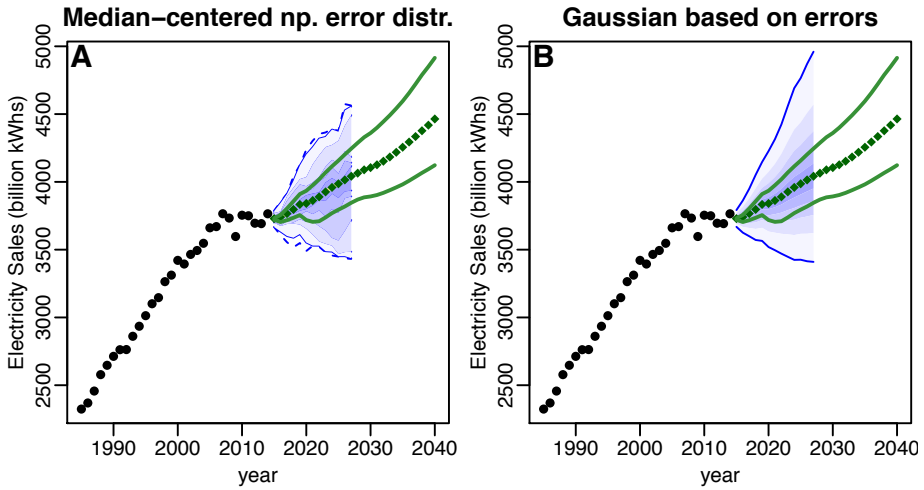

Figure 3: Density forecasts for electricity sales based on AEO 2016. (A) Median-centered non-parametric EPI (NP_2). The median or bias now coincides with the AEO reference case. (B) The Gaussian density forecast based on the SD of the errors (G_1), which was the best forecast over the test range. The envelope scenarios are narrower in both cases.

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*).

These two non-parametric estimations are compared to two parametric distributions, a Gaussian with a mean of zero and the variance of the errors (G_1) (Fig. 3) and with the variance of historical values (G_2) (Fig. 2). 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 [9]. 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 non-parametric error PDF. In addition, a Gaussian is much simpler to use as a model input. A discussion of normality and correlation in the errors is provided in the *SI*.

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

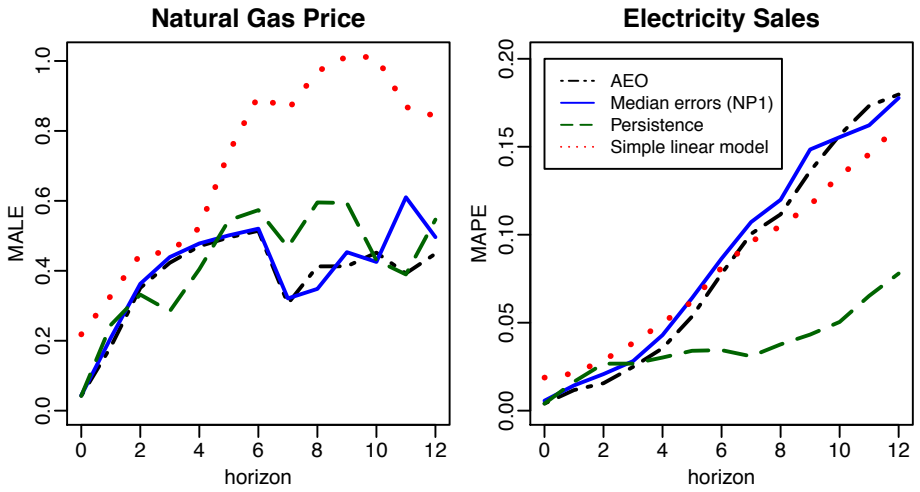

Figure 4: The mean absolute percentage or log-error (MAPE/MALE) for the test range 2003-2014. We see that for natural gas prices (in nominal \$), the median of NP_1 performs similarly to the AEO reference case. For electricity sales, the reference case outperforms the median for nearly every horizon.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

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 mean absolute percentage error (MAPE) and the mean absolute log error (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 2003-14 (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 has significantly outperformed the reference case for eight quantities. 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 has been 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 non-parametric uncertainty (NP_2) is advised for most quantities except those.

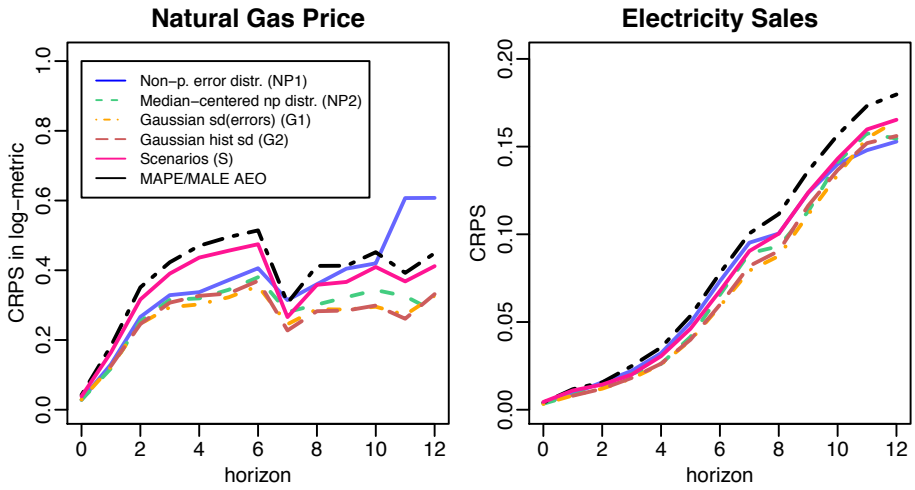

Figure 5: The continuous ranked probability score (CRPS) for the test range 2003-2014. A lower CRPS corresponds to a better density or ensemble forecast.

Gaussian density forecasts often perform well

Scoring rules, or scores, provide a mean for comparing the performance of different probabilistic forecasts. We use the continuous ranked probability score (CRPS), which is a strictly proper score in this case [26]. 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 [25], and allows for comparison with point and ensemble forecasts [26, 27] (*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. Comparing Fig. 4 and 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*).

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 nine out of the eighteen quantities and G_2 for three quantities. The performance of G_2 is however often similar to G_1 and it is second best for eight 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 if the empirical methods performed significantly better than uncertainty estimates based on the scenarios.

AEO scenario ranges are narrower than observed uncertainties

A varying number of scenarios, intended to give users insight about how the future might differ from the reference case under varied assumptions, are published in every AEO. No value is assigned to the probability that a future outcome will fall within or outside the scenario range. We are therefore not discussing a density forecast here but rather an ensemble forecast. The CRPS allows for comparison of a density forecast with an ensemble forecasts. It assigns every discrete scenario an equal point probability mass (method S). Because of the varying number of scenarios in the AEO, we make a simplification and only consider the reference case and the high and low envelope scenarios, which do not correspond to a specific scenario in the AEO. In addition, we compare to 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. This figure also includes the scores for SP_1 and SP_2. A normalized CRPS of an empirical method that is smaller than 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 in 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 were the better probabilistic forecasts 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*). The width

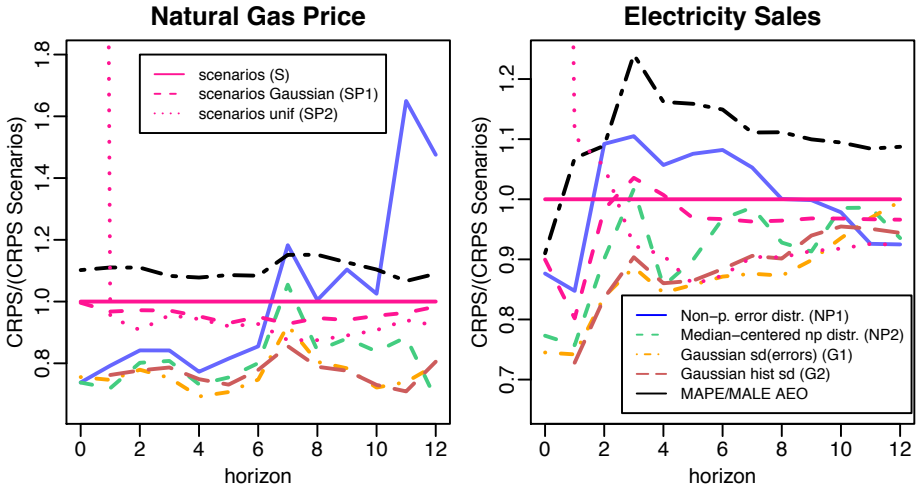

Figure 6: Relative improvement of the methods with respect to the envelope scenarios for the test range 2003-2014. Values are plotted as fraction of the CPRS of the scenario ensemble (S). A normalized CRPS lower than 1.0 corresponds to a better density forecast. SP_1 corresponds to a normal distribution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

between highest and lowest scenario, however, changes greatly from one AEO to another and is somewhat correlated to the number of scenarios published.

Discussion and Conclusion

There are empirical methods available for estimating the uncertainty around the AEO reference case, which have proven 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 standard deviation of forecast errors in their retrospective reports. We supply the values for the AEO 2016 in the *SI*. A non-parametric 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 non-stationary, there is no way to be sure that our results will be applicable to

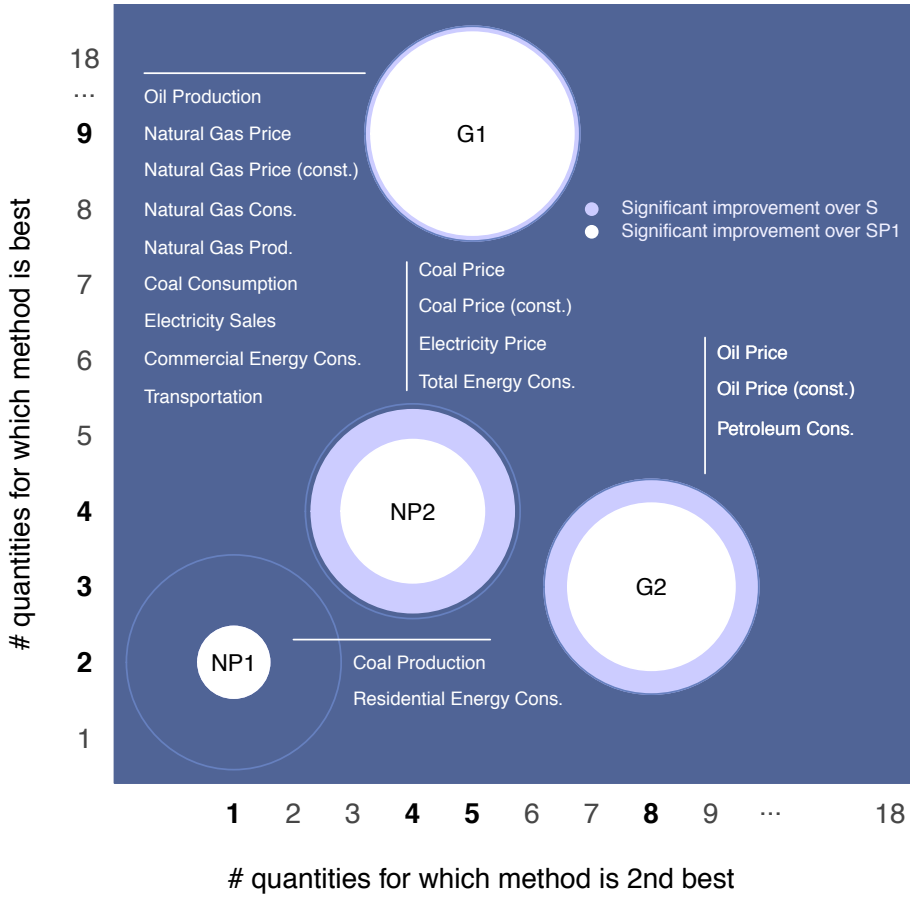

Figure 7: The Gaussian based on errors (G_1) performs best for the evaluation criteria, compared to the non-parametric biased EPI (NP_1), the non-parametric centered EPI (NP_2), and the Gaussian based on historical deviations (G_2). A better method is located in the upper part, and right. A larger circle corresponds to significant improvement over the scenarios for more quantities. The thin circle marks the max. radius of 18 quantities. Improvement is more likely over S (outer solid circle) than SP_1 (inner).

future data. However, the way we evaluated and 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 want to 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 AEO projections.

1 Materials and Methods

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

Data

The data set consists of AEOs 1982-2016 and historical values from 1985 to 2015. Historical data were taken from the EIA Retrospective Review [35] and the AEOs [34], and conversions were applied where necessary. All data are publicly available on the EIA website. Refer to *SI: Data Description* for more detail. The data analysis was performed in R [39].

List of methods

Point forecasting methods:

- *AEO reference case*: We treat the AEO reference case as a point forecast. The reference case is published as a projection of the current state of laws and regulations and does not represent a best estimate forecast. However, the reference case is the most consistent way to choose a best estimate.
- *Median errors (NP_1)*: The median of the EPI with a non-parametric distribution of the errors (NP_1), computed as the reference case adjusted by the median of the 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 7 historical observations. This size of window is the optimum for the test range.

Density forecasting methods:

- NP_1: EPI with a non-parametric distribution of the forecasting errors and a median different to the reference case. This method was originally published by [28].
- NP_2: EPI with a non-parametric 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: A Gaussian distribution with the standard deviation of the past errors and a mean and median of $\epsilon = 0$.
- G_2: Gaussian distribution with a standard deviation 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 projection in every year. These corresponds to the envelope of all scenarios by using only the highest and lowest projected values.
- SP: Two parametric density prediction based on the envelope scenarios in the AEO. We chose a Gaussian distribution with the distance to the farther scenario as 1 SD (SP_1) and a uniform distribution between the envelope scenarios (SP_2).

MAPE

The mean absolute percentage error (MAPE) is a measure for point forecast performance. This becomes the mean absolute log error (MALE) in the case of price forecasts with log-errors. They are defined as

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

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 in a sample for a particular horizon H. $\hat{y}$ refers to the forecast, while y is the actual observation.

CRPS

The continuous ranked probability score (CPRS) for every horizon, as we use it in this paper, is defined as

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

similar to [26]. ϵ_t is a point of the predictive error distribution, while ξ_t is the forecast error of the observation. The CRPS compares the 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 non-parametric CDF is computed like the CRPS for an ensemble forecast of discrete scenarios [27]. For ensemble forecasts, the CRPS can also be written as $CRPS_H(F,\epsilon) = \frac{1}{n_H}\sum_{t=1}^{n_H}\left[E_F\left|\epsilon_t - \xi_t\right| - \frac{1}{2}E_F\left|\epsilon_t - \epsilon_t'\right|\right]$ [26]. 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 $E_F\left|\epsilon_t - \xi_t\right| = \left|\xi_t\right|$ and $E_F\left|\epsilon_t - \epsilon_t'\right| = 0$. The CRPS is a strictly proper score here [26], 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 $\text{CRPS}_{S,H}$ of the scenario ensemble.

Improvement testing

We perform a bootstrap on the single CRPS results in a horizon sample, which then are used to compute the CRPS_H, and the aggregated CRPS average for the ranking. For every 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

For exploring the sensitivity of the ranking, we vary the default assumptions. Instead of first averaging the normalized CRPS and then rank that result, we alternatively first rank the CRPS_H and then averaged over the horizons. We also average over the full range of horizons $H = 1$ to $H = 12$ instead of the core range, that includes 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 performance 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.

2 Acknowledgements

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. E.D.S. 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 (EPRI) 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).

References

[1] J. J. Winebrake and D. Sakva, "An evaluation of errors in US energy forecasts: 1982–2003," *Energy Policy*, vol. 34, no. 18, pp. 3475–3483, 2006.

[2] M. Wara, D. Cullenward, and R. Teitelbaum, "Peak Electricity and the Clean Power Plan," *The Electricity Journal*, vol. 28, no. 4, pp. 18–27, 2015.

[3] A. Q. Gilbert and B. K. Sovacool, "Looking the wrong way: Bias, renewable electricity, and energy modelling in the United States," *Energy*, vol. 94, pp. 533–541, 2016.

[4] C. Fischer, E. Herrnstadt, and R. Morgenstern, "Understanding errors in EIA projections of energy demand," *Resource and Energy Economics*, vol. 31, no. 3, pp. 198–209, 2009.

[5] H. Linderoth, "Forecast errors in IEA-countries' energy consumption," *Energy Policy*, vol. 30, no. 1, pp. 53–61, 2002.

[6] V. Smil, "Perils of long-range energy forecasting: reflections on looking far ahead," *Technological Forecasting and Social Change*, vol. 65, no. 3, pp. 251–264, 2000.

[7] Intergovernmental Panel on Climate Change, "Definition of terms used within the DDC pages." `http://www.ipcc-data.org/guidelines/pages/definitions.html`, 2015.

[8] M. G. Morgan and D. W. Keith, "Improving the way we think about projecting future energy use and emissions of carbon dioxide," *Climatic Change*, vol. 90, no. 3, pp. 189–215, 2008.

[9] A. I. Shlyakhter, D. M. Kammen, C. L. Broido, and R. Wilson, "Quantifying the credibility of energy projections from trends in past data: The US energy sector," *Energy Policy*, vol. 22, no. 2, pp. 119–130, 1994.

[10] P. P. Craig, A. Gadgil, and J. G. Koomey, "What can history teach us? A retrospective examination of long-term energy forecasts for the United States," *Annual Review of Energy and the Environment*, vol. 27, no. 1, pp. 83–118, 2002.

[11] T. Gneiting, "Editorial: probabilistic forecasting," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 319–321, 2008.

[12] S. P. Vahey and L. Wakerly, "Moving towards probability forecasting," *BIS Paper*, no. 70b, 2013.

[13] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, pp. 125–151, 2014.

[14] F. X. Diebold, A. S. Tay, and K. F. Wallis, "Evaluating density forecasts of inflation: The survey of professional forecasters," in *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger* (R. F. Engle and H. White, eds.), Oxford University Press, 1997.

[15] E. Britton, P. Fisher, and J. Whitley, "The Inflation Report projections: understanding the fan chart," *Bank of England Quarterly Bulletin*, vol. 38, no. 1, pp. 30–37, 1998.

[16] M. Blix and P. Sellin, "Uncertainty bands for inflation forecasts," *Sveriges Riksbank Working Paper Series*, vol. 65, 1998.

[17] A. S. Tay and K. F. Wallis, "Density forecasting: a survey," *Journal of forecasting*, vol. 19, no. 4, pp. 235–254, 2000.

[18] T. J. Linsmeier and N. D. Pearson, "Value at risk," *Financial Analysts Journal*, vol. 56, no. 2, pp. 47–67, 2000.

[19] A. E. Raftery, N. Li, H. Ševčíková, P. Gerland, and G. K. Heilig, "Bayesian probabilistic population projections for all countries," *Proceedings of the National Academy of Sciences*, vol. 109, no. 35, pp. 13915–13921, 2012.

[20] P. E. McSharry, S. Bouwman, and G. Bloemhof, "Probabilistic forecasts of the magnitude and timing of peak electricity demand," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 1166–1172, 2005.

[21] J. W. Taylor, P. E. McSharry, and R. Buizza, "Wind power density forecasting using ensemble predictions and time series models," *IEEE Transactions on Energy Conversion*, vol. 24, no. 3, pp. 775–782, 2009.

[22] P. Pinson, "Wind energy: Forecasting challenges for its operational management," *Statistical Science*, vol. 28, no. 4, pp. 564–585, 2013.

[23] F. X. Diebold, T. A. Gunther, and A. S. Tay, "Evaluating density forecasts, with applications to financial risk management," *International Economic Review*, vol. 39, pp. 863–883, 1998.

[24] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 243–268, 2007.

[25] L. A. Smith, E. B. Suckling, E. L. Thompson, T. Maynard, and H. Du, "Towards improving the framework for probabilistic forecast evaluation," *Climatic Change*, vol. 132, no. 1, pp. 31–45, 2015.

[26] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[27] H. Hersbach, "Decomposition of the continuous ranked probability score for ensemble prediction systems," *Weather and Forecasting*, vol. 15, no. 5, pp. 559–570, 2000.

[28] W. H. Williams and M. L. Goodman, "A simple method for the construction of empirical confidence limits for economic forecasts," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 752–754, 1971.

[29] P. Pinson and G. Kariniotakis, "Conditional prediction intervals of wind power generation," *Power Systems, IEEE Transactions on*, vol. 25, no. 4, pp. 1845–1856, 2010.

[30] NOAA National Hurricane Center, "National hurricane center forecast verification," March 2016. `http://www.nhc.noaa.gov/verification/verify6.shtml`.

[31] O. Isengildina-Massa, S. Irwin, D. L. Good, and L. Massa, "Empirical confidence intervals for usda commodity price forecasts," *Applied Economics*, vol. 43, no. 26, pp. 3789–3803, 2011.

[32] M. Knüppel, "Efficient estimation of forecast uncertainty based on recent forecast errors," *International Journal of Forecasting*, vol. 30, no. 2, pp. 257–267, 2014.

[33] Y. S. Lee and S. Scholtes, "Empirical prediction intervals revisited," *International Journal of Forecasting*, vol. 30, no. 2, pp. 217–234, 2014.

[34] U.S. Energy Information Administration, "Annual Energy Outlook," 2016. `http://www.eia.gov/forecasts/aeo/`.

[35] U.S. Energy Information Administration, "Annual Energy Outlook Retrospective Review," 2015. `https://www.eia.gov/forecasts/aeo/retrospective/`.

[36] B. C. O'Neill and M. Desai, "Accuracy of past projections of us energy consumption," *Energy Policy*, vol. 33, no. 8, pp. 979–993, 2005.

[37] M. Auffhammer, "The rationality of eia forecasts under symmetric and asymmetric loss," *Resource and Energy Economics*, vol. 29, no. 2, pp. 102–121, 2007.

[38] C. M. Sprenkle, "Warrant prices as indicators of expectations and preferences," *Yale Economics Essays*, vol. 1, pp. 178–231, 1961.

[39] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[40] N. M. Razali and Y. B. Wah, "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011.