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

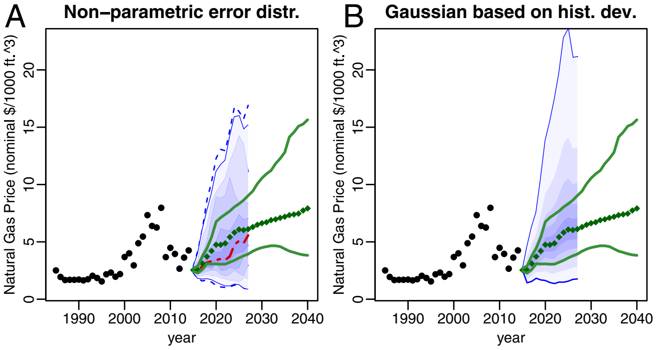

Fig. 1. Density forecasts for natural gas prices in nominal dollars. (*A*) Nonparametric EPI based on forecast errors (NP_1). (*B*) Gaussian density forecast based on the variability of historical values (G_2), which tested to be the better estimate. Historical values are indicated by black circles, 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.

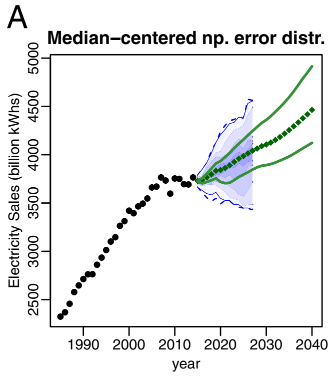

Fig. 2. Density forecasts for electricity sales based on AEO 2016. (*A*) For the median-centered nonparametric EPI (NP_2), the median or bias now coincides with the AEO reference case. (*B*) The Gaussian density forecast based on the SD of the errors (G_1) was the best forecast over the test range. The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

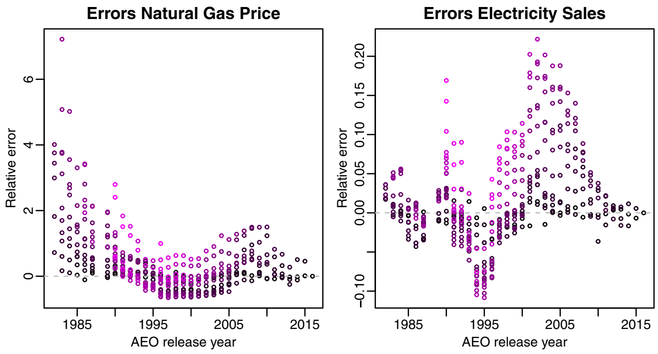

Fig. 3. Forecast errors by AEO release year. Different colors correspond to forecast horizons ranging from $H = 0$ in black to $H = 21$ in purple. All forecast errors are untransformed. Note the different scale. No AEO was released for 1988.

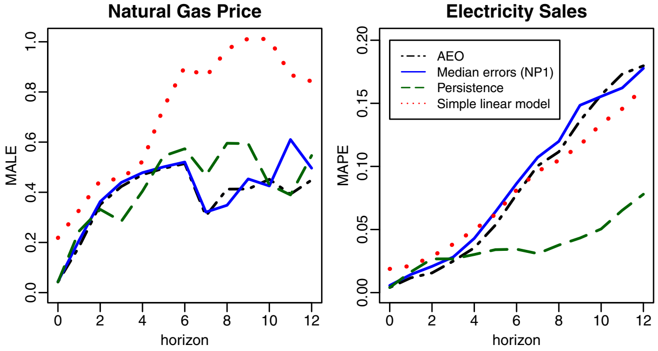

Fig. 4. The mean absolute percentage or log error (MAPE/MALE) for the test range 2003–2014. We see that for natural gas prices (in nominal dollars), the median of NP_1 performs similarly to the AEO reference case. For electricity sales, the reference case outperforms the median for nearly every horizon. 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

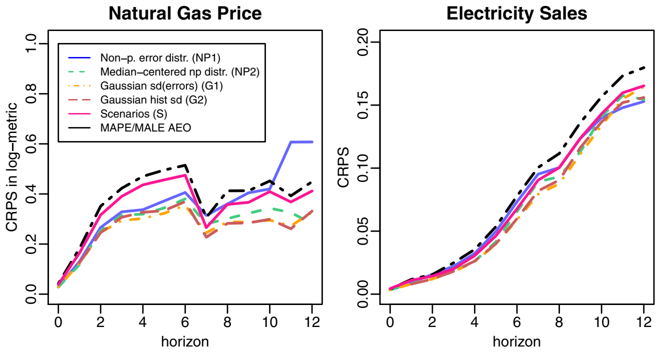

Fig. 5. The CRPS for the test range 2003–2014. A lower CRPS corresponds to a better density or ensemble forecast.

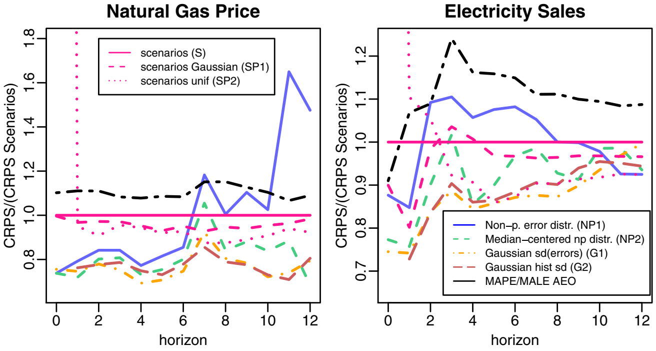

Fig. 6. Relative improvement of the methods with respect to the envelope scenarios for the test range 2003–2014. Values are plotted as fraction of the CPRS of the scenario ensemble (S). A normalized CRPS lower than 1.0 corresponds to a better density forecast. SP_1 corresponds to a normal distribution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

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

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

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

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 $CRPS_H(F, \epsilon) = \frac{1}{n_H} \sum_{t=1}^{n_H} \left[E_F \,|\epsilon_t - \xi_t| - \frac{1}{2} E_F \left|\epsilon_t - {\epsilon'}_t\right|\right]$ (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 $E_F \,|\epsilon_t - \xi_t| = |\xi_t|$ 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

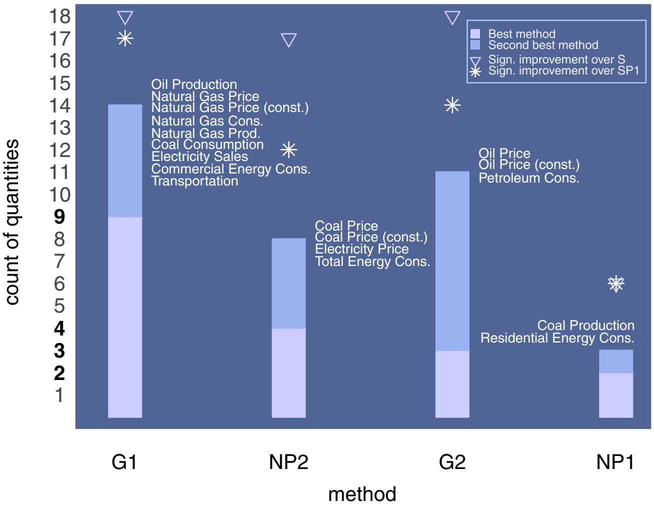

Fig. 7. Graphical summary of the evaluation results. The methods are ordered by the number of quantities they perform best for (listed in white). The Gaussian based on errors (G_1) performs best or second best for 14 of 18 and showed significant improvement over the scenarios for almost all quantities. 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).