

## Perspective

## The need for better statistical testing in data-driven energy technology modeling

C. Lennart Baumgärtner,<sup>1,2,\*</sup> Rupert Way,<sup>1,2</sup> Matthew C. Ives,<sup>1,2</sup> and J. Doyne Farmer<sup>1,2,3,4</sup>

## SUMMARY

Technology modeling is a vital part of developing and understanding energy system scenarios and policy, but it is challenging due to data limitations, deep uncertainty, and the complex social and technological dynamics involved in the evolution of energy systems. These difficulties are often compounded by unsound technology forecasting practice, including overfitting, data selection bias, and *ad hoc* assumptions, leading to unreliable conclusions. We flag several cases where this has been problematic and analyze in detail a recent model for predicting the pace of solar photovoltaic and wind energy deployment. We discuss general takeaways and provide suggestions for how statistical testing should be conducted to avoid such problems in the future and to quantify the reliability of forecasts.

## INTRODUCTION

To decarbonize our energy system, we must dramatically change the technologies we use to produce, store, and utilize energy. To do this, we need to cope with uncertainty about future technology costs and any potential limits to their deployment. In particular, to make sound investments and effective public policy, we require reliable technology forecasting models whose accuracy can be quantified using standard statistical methods.<sup>1</sup> Unfortunately, such models are all too often built without proper statistical testing and with insufficient effort made to understand the reliability of their conclusions.

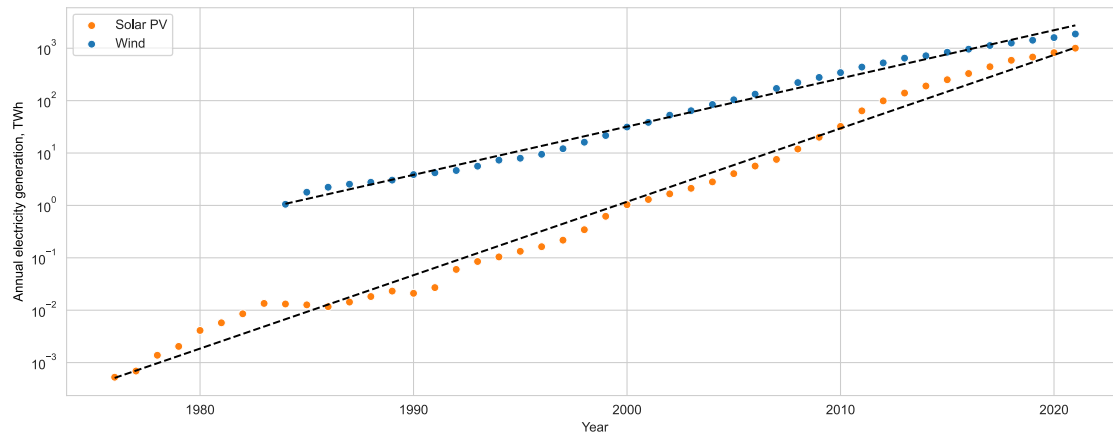
One symptom of this problem is the tendency to refer to the outputs of technological change models as “projections.” This word is usually invoked to signal a lack of confidence in the predictive power of the model. Nonetheless, there are many circumstances where projections are used to inform wider energy system scenarios or policy decisions, meaning that they are *de facto* conditional forecasts about cause and effect, even if they are not labeled as such. Such forecasts should be subjected to the same statistical testing and validation procedures that are routine in disciplines such as finance, economics, epidemiology, and meteorology. Without statistical testing, it is impossible to know how reliable we should expect forecasts to be.

Here, we focus on data-driven, statistically based technological change models that generate forecasts using historical deployment data, historical cost data, or some combination of the two. We will not address equilibrium models (such as most integrated assessment models) or bottom-up engineering models.<sup>2</sup> Because data-driven technological change models are often used both to calibrate large energy system models and to inform the wider discourse on scenario feasibility, proper statistical testing and validation are critical.

## CONTEXT &amp; SCALE

Forecasting the future evolution of clean energy technologies is vital to the energy transition. To develop effective policies and make sound investments, energy technology models need to be reliable, and we need scientifically justified ways to assess their reliability. We analyze the reliability of several data-driven energy technology models from the recent literature, including several “S-curve” models, which have recently gained attention in modeling technology diffusion. Our examination shows that, due to a lack of statistical testing, many such models produce unreliable results, for example, underestimating the future deployment of solar photovoltaics and overestimating the future cost of batteries. We highlight the importance of statistical testing and describe various methods to validate a model's reliability.





**Figure 1. Global electricity generation from wind (onshore and offshore) and solar PV**

The historical annual generation of electricity from wind and solar PV shows approximately exponential growth.<sup>5,6</sup>

We begin by focusing on the deployment of solar photovoltaic (PV) and wind, which will be the main topic in this paper. Solar PV and wind play an important role in most national and global future energy system scenarios, such as those presented in the Intergovernmental Panel on Climate Change's (IPCC's) most recent assessment report, AR6.<sup>3</sup> What insights can historical data provide about the likelihood of meeting the deployment levels in different scenarios and the cost of achieving such scenarios? As shown in Figure 1, over the last 35 years, wind electricity production has grown at an average rate of 21% per year, and over the last 45 years, solar PV has grown at an average rate of 32% per year. These rapid growth rates will necessarily diminish through time and eventually level off, but how and when will this happen? Their future deployment rates will most likely be a key determinant of whether or not we can meet the Paris Agreement,<sup>4</sup> and predictions about this are essential for planning to meet these goals.

Technology deployment trends are frequently characterized by a standard family of models called S-curves. In the early stage of development, absolute deployment levels are small, but growth is rapid. As a technology matures, its growth slows and flattens until it reaches market saturation. This inevitably happens due to technology competition, consumer preferences, and other constraints. For solar PV and wind, slower growth could also occur for socio-political reasons, such as land-use restrictions, public opposition, incumbent interests, insufficient investment, inadequate policy and institutional support, or a lack of cheap energy storage technologies to manage intermittency. However, although it is easy to imagine technology-specific barriers such as these, it is often hard to predict our ability to find solutions, and there is no reliable evidence regarding the timing or size of potential impacts (for example, until recently, grid-scale electricity storage was considered a severe constraint on variable renewables, but this is changing due to plummeting battery costs).

There are many functional forms that can be used to represent S-curves, the simplest example of which is the logistic function. It is written

$$x_t = S(t) = \frac{L}{1 + e^{-k(t - t_0)}}, \quad (\text{Equation 1})$$

where  $t$  is time in years,  $x_t$  is the number of units of the technology deployed at time  $t$ ,  $L$  is the asymptotic diffusion level,  $k$  is the initial annual exponential growth rate, and  $t_0$  is a location parameter centering the S-curve in time. In the limit  $t \ll t_0$ ,

<sup>1</sup>Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford OX1 3UQ, UK

<sup>2</sup>Smith School of Enterprise and the Environment, University of Oxford, Oxford OX1 3QY, UK

<sup>3</sup>Macrocosm Inc, 235 E. 4th St., Brooklyn, NY 11218, USA

<sup>4</sup>Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

\*Correspondence:  
lennart.baumgaertner@worc.ox.ac.uk  
<https://doi.org/10.1016/j.joule.2024.07.016>

$S(t) \approx Le^{k(t-t_0)}$ , and in the limit  $t \gg t_0$ ,  $S(t) \approx L$ . The growth rate  $dx_t/dt$  initially increases exponentially, then begins to noticeably slow down after reaching its maximum at the inflection point  $t = t_0$ , and decreases dramatically as  $x_t$  approaches its asymptotic value  $L$ . This functional form has been shown to provide a good representation of observed data in a variety of technologies, from large infrastructure projects (roads, canals, etc.) to consumer goods (mobile phones, internet users, etc.),<sup>7</sup> and we expect energy technologies to follow the same pattern. However, although S-curves often provide a good fit to observed data after a technology has already reached global saturation, real-world data never follows an S-curve exactly, and as yet, S-curve based methods have not been tested to assess their performance as forecasting tools for technologies in general.

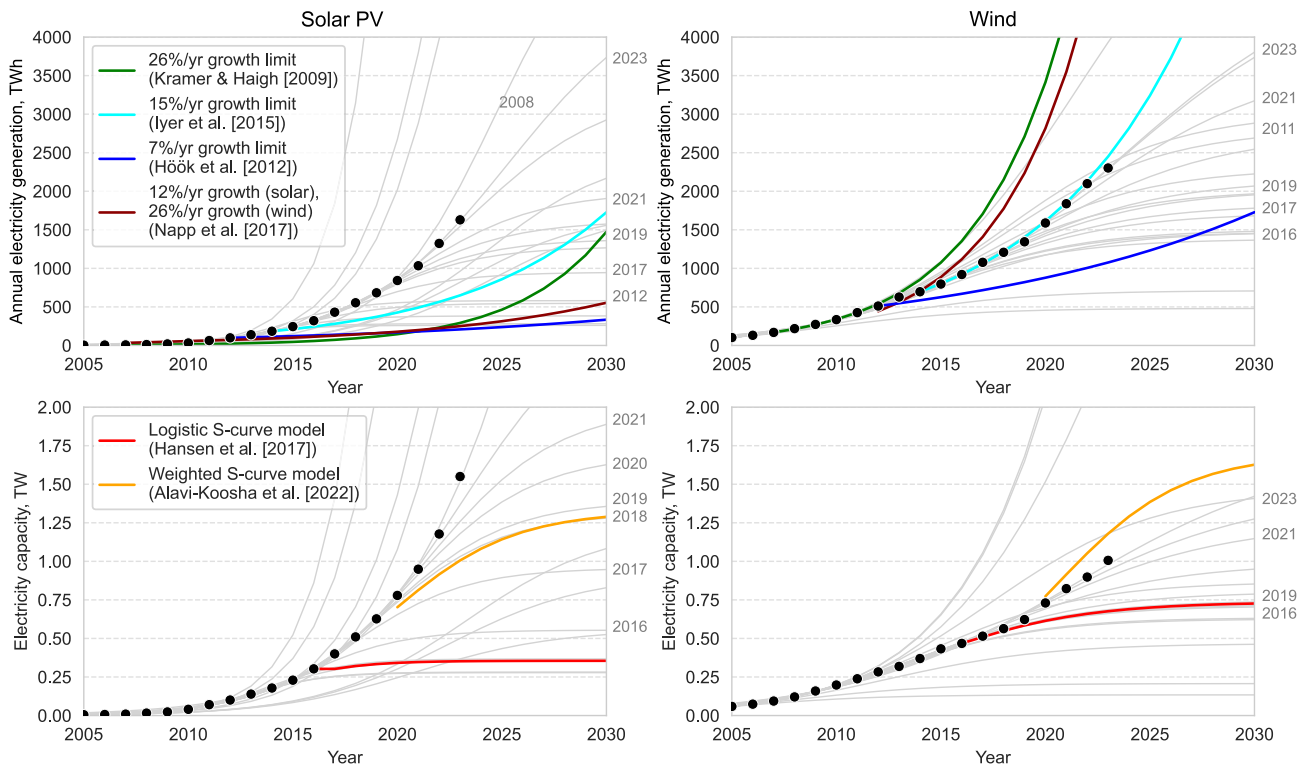
Modeling the future diffusion of novel low-carbon technologies is challenging, given the complex social and technological dynamics associated with the energy system.<sup>8,9</sup> The pace of change means that the clean technologies we are most interested in have limited historical records, and their socio-economic environment is rapidly evolving. This poses an issue for all forward-looking models, especially for data-driven approaches relying on historical observations for model calibration and validation. However, this does not mean that S-curve models are useless in this context; rather, it means that the reliability of their predictions needs to be assessed carefully so that we know how useful they are. As discussed later, for technologies such as solar PV and wind that are still in the early stages of their development, due to data limitations, it is necessary to test any new prediction methodologies on many different technologies, particularly those that have already reached maturity. This provides ground truth for testing predictions in situations where the correct answers are known.

Data limitations mean that it is very easy to overfit historical data, i.e., to choose models that match the noisy features of the data rather than the underlying signal. This typically occurs when the data are noisy, and the number of data points is small compared with the number of parameters. Other important problems are selection bias, where the selective exclusion of data distorts results, and the use of unfounded *ad hoc* assumptions that are chosen for convenience rather than being based on evidence (see Petropoulos et al.<sup>10</sup> for a comprehensive overview of data-driven predictive models and their challenges). Unfortunately, these mistakes are often made in statistical technological change models, resulting in poor predictions.

In this paper, we primarily focus on the problem of forecasting technology deployment and mention several examples that illustrate such mistakes. We also provide an example from cost forecasting to demonstrate that such problems are not unique to the prediction of deployment. We begin by discussing a range of examples where predictions were unreliable in hindsight, then dive deeply into one specific example as a case study. We conclude by discussing the issues of overfitting, selection bias, and poor statistical practice more generally, as well as providing some guidelines about making more reliable models and assessing the quality of their predictions.

## EXAMPLES OF OVERFITTING AND SELECTION BIAS

The difficulty of forecasting technology deployment directly using the logistic function is illustrated in [Figure 2](#), which shows point forecasts for the deployment of solar PV (left) and wind (right), in terms of annual electricity generation (top) and generating capacity (bottom). We pretend to be at a time  $t_{\max}$  in the past, fitting logistic S-curves on all data prior to  $t_{\max}$ . The gray lines in [Figure 2](#) show logistic S-curve



**Figure 2. Global deployment for solar and wind (onshore plus offshore) vs. past deployment projections**

Annual electricity generation (top row) and generating capacity (bottom row) for solar (left) and wind (right). (Dots) Actual annual electricity generation and capacity. Data prior to 2005 is not shown here for better readability. (Gray lines) S-curve fits using the logistic function described in the main text, for  $t_{\max}$  between 2005 and 2019. We see that S-curves are highly sensitive to  $t_{\max}$ , as indicated by the gray labels. Recent fits align more closely with the data; however, it is unknown how well they will align with future deployment. (Colorful lines) Predictions for future solar and wind deployment in the literature. Similar to the S-curve fits, future deployment predictions vary widely in the literature. For solar, all predictions underestimate future deployment. Data until 2019 based on IEA WEB,<sup>11</sup> until 2023 based on Ember.<sup>12</sup>

fits for different values of  $t_{\max}$ . As the figure illustrates, the fits are highly sensitive to  $t_{\max}$ . The fits for solar vary from predicted asymptotic capacities as low as a quarter of a terawatt (a factor of six lower than today's capacity) to extremely high values that are too large to be seen in the plot. In later years, the fits are less sensitive to  $t_{\max}$  but remain unstable. Furthermore, the fits in recent years tend to underestimate future diffusion, as is evident by the fact that most of them are to the right of the real data.

The inconsistent S-curve fits in Figure 2 provide a good example of overfitting. Although the logistic function may seem simple, it is a nonlinear function with three free parameters, and even small amounts of noise can make fits unstable until diffusion is sufficiently developed, which is not true yet for either solar or wind. Despite this, S-curve fits have been used multiple times in the literature to forecast future renewables growth. For example, Hansen et al.<sup>13</sup> (Figure 2, red line) fitted S-curves to data up until 2015 and concluded that “the logistic model implies that the total installed [wind and solar PV] capacity saturates at around 1.8 TW in 2030.” Although these claims have since been refuted methodologically,<sup>14</sup> and the fact that by 2023 we had already installed 2.4 TW of solar PV and wind,<sup>15</sup> such statements are still persistent in the literature (e.g., Madsen and Hansen<sup>16</sup>).

Another example of likely overfitting is given in Alavi-Koosha et al.,<sup>17</sup> who forecasted future renewable capacities using a weighted average of four different kinds of



S-curves (Figure 2, orange line). Although weighting different models can lead to better forecasts, it is not a panacea—if all the forecasts are bad or if there are systematic biases, the resulting forecasts may still be bad. As shown in Figure 2, the forecasts have been significantly too low for solar PV and too high for wind, indicating that the underlying model is likely overfit.

Perhaps more common than overfitting is the issue of selection bias. We have found a number of examples in the literature that have analyzed historical growth rates in (energy) technologies to imply limits for future renewable growth rates.

In 2009, Kramer and Haigh<sup>18</sup> (Figure 2, green lines) used historical data from past energy technologies to argue that annual growth in clean energy technologies is likely to be at most 26% during their early deployment phase, until they achieve “materiality” (specified as “around 1%” of global primary energy production), after which annual growth will slow dramatically and proceed linearly. However, if we assume 26% annual growth from 2009, materiality will not happen before 2030 for solar PV or before 2017 for wind. As shown in Figure 2, these limits provide a reasonable upper bound for wind but have already been surpassed by solar PV.

Höök et al.<sup>19</sup> also studied growth rates in previous energy technologies (nuclear, liquid natural gas, biofuels, and wind) and concluded that wind and solar PV electricity generation are unlikely to grow faster than 7% per year (Figure 2, blue line). This is based on the fact that fossil fuels have never grown faster than this rate over multiple decades. This comparison was repeated in a similar fashion by Smil in 2016,<sup>20</sup> although prior renewable growth had already been significantly larger than 7%.

Iyer et al.<sup>21</sup> built on the work of Kramer and Haigh<sup>18</sup> and Höök et al.,<sup>19</sup> to examine the historic deployment rates of around 40 technologies, including non-energy technologies, such as washing detergent and black-and-white televisions. They concluded based on their wider subset of technologies that annual growth rates above 15% were unlikely for renewables and applied this limit to an integrated assessment model (GCAM) (Figure 2, cyan lines). Although adequate for wind electricity, this also under-estimated future growth in solar PV.

Napp et al.<sup>22</sup> performed a similar analysis, limiting renewable growth to 20% by the same arguments. They applied these limits to another integrated assessment model (TIAM-Grantham) (Figure 2, brown lines). Again, this upper bound might be reasonable for wind but substantially underestimates growth in solar PV.

In all of these examples, researchers investigated a sample of historic technology growth trajectories and then inferred limits on future renewable growth. Although mostly adequate for wind, all of these limits were too low for solar PV, as evident from Figure 2. It appears that these studies were all subject to selection bias; their conclusions are biased because too few technologies were considered in their analyses. For example, mobile phone and internet users both grew at rates up to 30% annually, comparable to solar PV. We discuss these high growth rate examples in more detail in the next section.

Issues with overfitting and data selection bias also occur when modeling the cost of technologies. For example, in 2019, Hsieh et al.<sup>23</sup> applied a two-stage learning model to lithium-ion battery prices, in which raw material prices were treated separately from material synthesis and battery pack manufacturing costs. They modeled

raw material prices of lithium, cobalt, and nickel by extrapolating from a very limited selection of the historical price record and predicted that in 2023 nickel-manganese-cobalt (NMC) battery pack prices would be in the range 170–201 USD(2023)/kWh. However, cobalt prices have dropped significantly since their publication, and in 2023, the average NMC battery pack price was at an all-time low of 146 USD(2023)/kWh.<sup>24,25</sup> The average pack price for battery electric vehicles was lower, at 128 USD(2023)/kWh,<sup>24</sup> due to a shift toward lower cost lithium-iron-phosphate batteries, and all pack prices have continued to decline in 2024.<sup>26</sup> This indicates that their forecasts were likely overfit. Subsequently, Penisa et al.<sup>27</sup> tested a similar learning model for predicting EV battery costs and rejected a model that includes lithium and cobalt metal prices on the basis that it was overfit. Nagy et al.<sup>28</sup> showed that multi-factor learning models tend to overfit data because they have too many parameters and therefore provide worse forecasts of future costs than simple alternatives. The approach developed by Hsieh et al.<sup>23</sup> is nevertheless being used to forecast battery prices in the US decades into the future.<sup>29</sup>

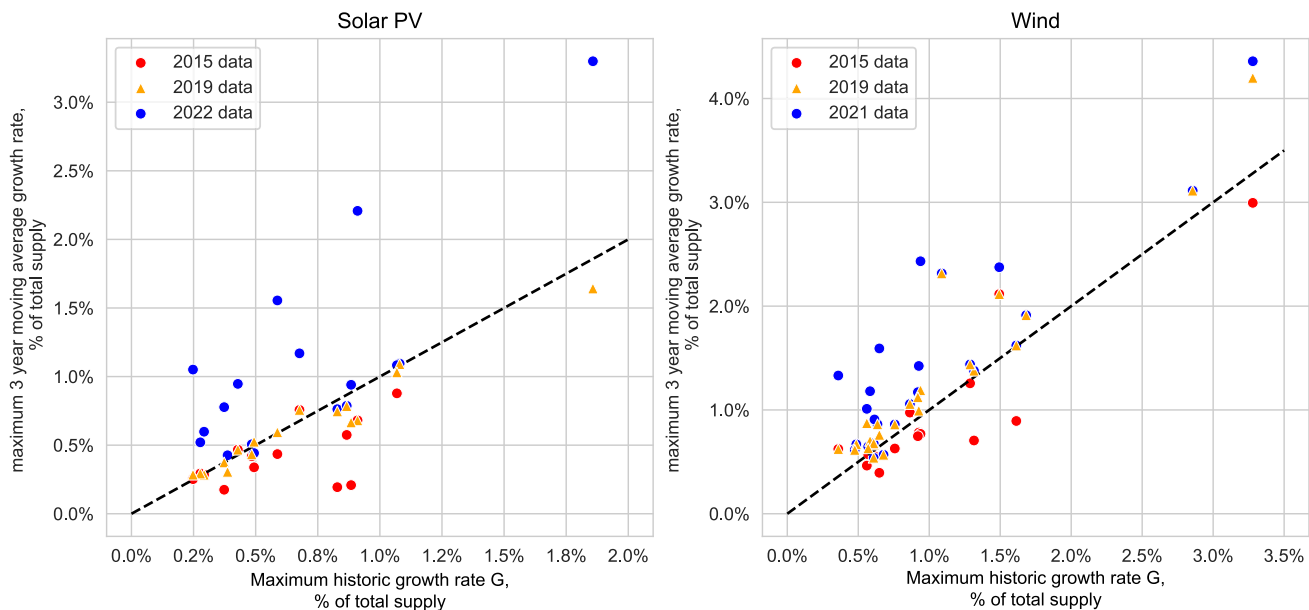
## TESTING AN ENERGY TECHNOLOGY DIFFUSION MODEL FROM THE RECENT LITERATURE

To highlight how overfitting and selection bias can affect the reliability of a statistical technological change model, we consider in detail a recent example from the literature. Cherp et al.<sup>30</sup> developed a model of deployment growth in solar PV and wind to assess the feasibility of rapid deployment in energy scenarios. Based on historical growth in these two technologies, they conclude that “some 1.5°C and 2°C pathways pose serious feasibility concerns” and “replicating or exceeding the fastest national (historic) growth may be challenging.”

These conclusions have attracted significant attention in the academic community and wider media,<sup>31</sup> and their methodology has been replicated to comment on feasible national trajectories for nuclear energy<sup>32</sup> and carbon capture and storage (CCS).<sup>33</sup> In the following, we discuss their method and explain why it suffers from problems of overfitting and selection bias.

The method of Cherp et al.<sup>30</sup> aims to take advantage of the fact that some countries already have substantial deployment of solar and wind energy, whereas others have little to none. They fit S-curves to historical diffusion data in 60 countries to estimate what they call the maximum historic growth rate  $G$  for each technology in each country.  $G$  is calculated as the maximum growth rate of an S-curve fitted to the annual electricity production time series of a given country, measured relative to that country's total electricity supply. ( $G$  should not be confused with the annual exponential growth rates quoted earlier.)

The maximum growth rate  $G$  of the logistic S-curve occurs at the inflection point  $t_0$ , where  $G = kL/4$  (normalized by the total electricity supply at  $t_0$ ). Based on logistic S-curve fits, Cherp et al.<sup>30</sup> identify what they believe are countries that have already passed their period of maximum growth. This is done by comparing the current deployment level  $x_t$  with the fitted asymptotic market size  $L$ . For a given country at time  $t$ , they deduce that it has already passed its period of maximum growth if the maturity  $m = x_t/L > 0.5$ . (The value 0.5 is a natural choice as it corresponds to the diffusion level relative to the asymptote  $L$  at the inflection point  $t_0$ .) Cherp et al. then compare their estimates of  $G$  for these “high maturity” countries with the growth rates required for solar and wind in rapid decarbonization scenarios. (See [Section S1.1](#) for a detailed description of the method.) Because their estimates



**Figure 3. Maximum historic growth rates  $G$  for solar photovoltaics (left) and onshore wind energy (right) vs. maximum 3-year moving average growth rates until 2015 (red dots), 2019 (orange triangles), and 2022/2021 (blue dots)**

For a given color, each dot represents a different country. The observed growth rates for each country as a percentage of its annual electricity supply are calculated as the maximum 3-year moving average of the annual growth rate (Extended Data Figure 7B in Cherp et al.<sup>30</sup>) and plotted against the maximum historic growth rates  $G$  as calculated through S-curve fits performed by Cherp et al. For solar PV, the red dots, orange triangles, and blue dots consider the full national time series until 2015, 2019, and 2022, respectively. For onshore wind, the red dots, orange triangles, and blue dots consider the time series until 2015, 2019, and 2021, respectively, due to data availability. For countries with significant electricity generation by offshore wind, we used data until 2020 because the 2021 data were not yet available (the sample size has dropped to 17 countries for solar PV and 27 countries for onshore wind because a majority of countries are excluded by the filtering method of Cherp et al.). The 2019 data that were used to produce the fitted values of  $G$  are consistent with their predictions, but later out-of-sample data substantially exceed the predicted maximum rates. Similarly, earlier data fall below the predicted maximum rates. Data for solar are from Ember,<sup>12</sup> and data for wind are from IRENA<sup>6</sup> and IEA.<sup>11</sup>

of the maximum growth rates are lower than those required in 1.5°C or 2°C scenarios, they conclude that reaching our climate goals will be “challenging.”

Overfitting occurs when data are well matched by a model in sample but poorly matched out of sample. Here, “in-sample” refers to data used to calibrate a model, meaning to estimate its parameters or perform model selection, and “out-of-sample” refers to data that were not used for calibration. Because Cherp et al.<sup>30</sup> only analyzed data up to and including 2019, we can test for overfitting by using more recent data. Figure 3 compares actual with predicted maximum growth rates for individual countries in and out of sample. Rather than relying on an S-curve fit, we directly estimate the maximum growth rate observed at any point in time by taking the maximum average growth rate over all possible historical 3-year intervals. (3-year intervals are chosen to reduce noise.) We then compare our direct measurements with the maximum historical growth rates estimated by Cherp et al. When doing this, we only consider countries that Cherp et al. believe to have already reached or passed their point of maximum growth. The out-of-sample comparison is based on data through 2022 for solar PV and 2021 for onshore wind.

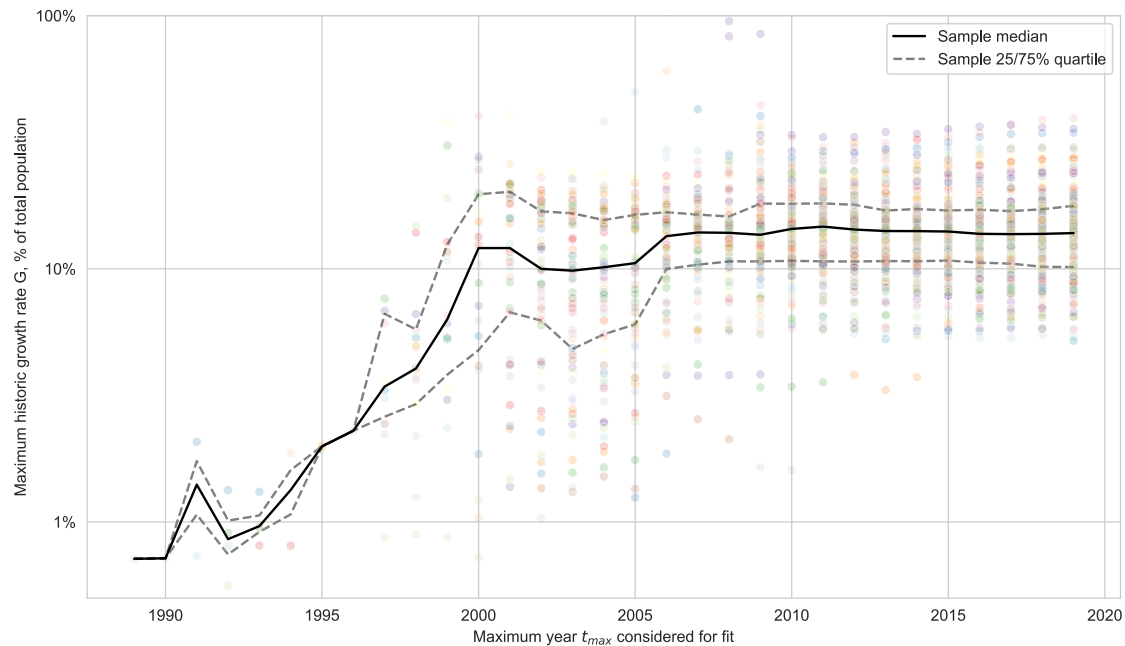
Figure 3 plots the growth rates observed in each year against those predicted by Cherp et al.<sup>30</sup> The in-sample growth rates in 2019, the year of their study, are clustered along the identity line, as they should be. This is not the case, however, for later out-of-sample growth rates. All but three of the out-of-sample data points (shown as blue dots) are above the identity line, and many are substantially above it. Although

Cherp et al. predict that growth rates above 2% for solar and 3.5% for wind should be rare, the observed maximum growth rates are as high as 3.3% for solar and 4.4% for wind. This is a clear demonstration of overfitting.

As mentioned in the introduction, in situations where the data are severely limited, such as for solar PV and wind diffusion, it is essential to perform tests on other technologies. This is also required to assess whether model dynamics are likely to be stationary and, if not, what the effects might be. The S-curve model is technology agnostic, meaning that all technologies follow some form of an S-curve, and it can be applied to any technology.<sup>7</sup> As with the logistic S-curve model, there is nothing about the method by Cherp et al.<sup>30</sup> that is specific to solar or wind energy. This can be used to vastly enlarge the quantity of data available for testing. Indeed, pooled technology datasets have already been extensively used in developing cost forecasting models.<sup>28,34,35</sup> We take advantage of the fact that the method of Cherp et al. is technology agnostic to test it on the global level and get more insight into its reliability. For this purpose, we choose mature technologies that are already close to asymptotic deployment and therefore provide ground truth. Thus, unlike solar and wind energy, which are still at a stage where their asymptotic levels of deployment are unknown, we can apply their method during different stages of a technology's development and then compare the results with what really happened. In the main text, we present the results of a back-test for mobile phone cellular subscriptions,<sup>36</sup> and in [Section S1.3](#), we present the same analysis for annual airline passenger miles and internet users, which all provide similar results.

A visual examination of the mobile phone data in [Figure S2](#) makes it clear that by 2020, the technology diffusion process for mobile cellular subscriptions was well past its inflection point, with the S-curve stagnating at the global and most national levels. Fitting a logistic S-curve to the global data yields a maximum growth rate  $G$  of roughly 9% at the inflection point  $t_0 \approx 2008$  and an asymptote  $L \approx 123\%$  of the total population. The estimated initial growth rate per year for mobile phones is  $k \approx 30\%$ , which is comparable to the annual growth rates for solar and wind energy. In the same way that Cherp et al.<sup>30</sup> normalize solar and wind deployment to total electricity supply, we normalize the number of phone subscriptions to the national (global) population. We use their code and data selection procedure to calculate the distribution of maximum national historical growth rates  $G$  for all possible maximum years  $t_{\max}$ .

The change in the distribution over time is shown in [Figure 4](#). The dots are the national maximum growth rates  $G$  for countries with  $m > 0.5$ . The solid line shows the median of the resulting distribution, and dotted lines indicate interquartile ranges (25% and 75%). The figure makes it clear that the distribution of  $G$  strongly depends on the choice of  $t_{\max}$ . The change in the median is indicative of the change in the distribution as a whole. The distribution briefly appears (roughly) stationary with a median of about 1% in the early phase running from  $t_{\max} = 1989$  to  $t_{\max} = 1993$ . After  $t_{\max} \approx 2005$ , the distribution is roughly stationary again, but the median is more than a factor of 10 higher than it is in the early stage. The change happens abruptly: When  $t_{\max} = 1995$ , the median of  $G$  is below 2%; when  $t_{\max} = 1998$ , it is about 4%, and when  $t_{\max} = 2000$ , it is about 10%, which is still only around 75% of the final value of 13.3%. (The fact that this is higher than the estimated growth rate of 9% above is due to variations in the S-curve fits with different final years.) Prior to reaching global maturity, the three technologies that we test all exhibit non-stationary behavior in the distribution of  $G$ , only stabilizing once they are close to maturity. This indicates that, once the dust settles, it is likely that we will see similar results for wind and solar.

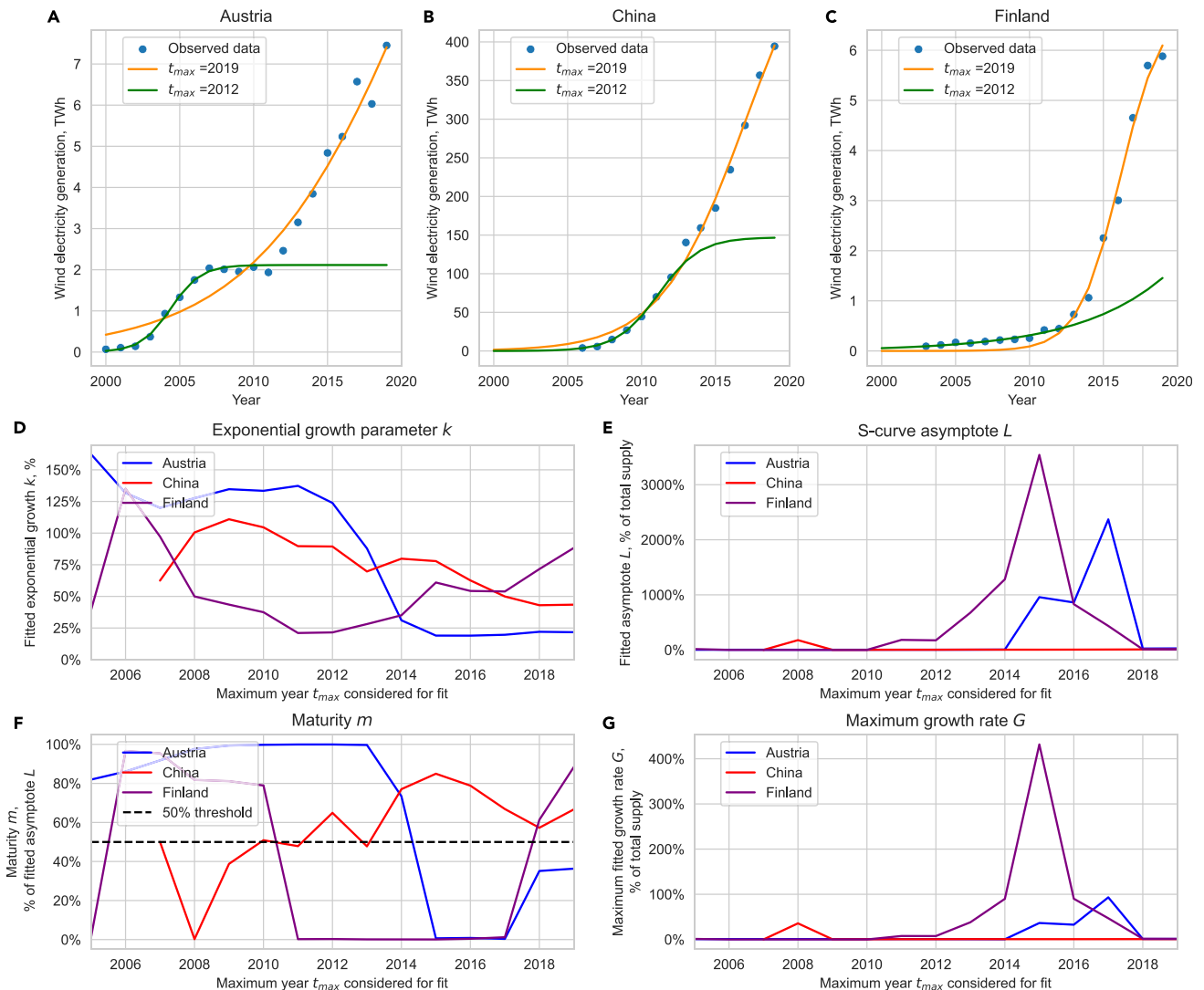


**Figure 4. Distribution of national maximum historic growth rates ( $G$ ) for mobile phones vs. the maximum year considered for the S-curve fit ( $t_{\max}$ )**  
National maximum historic growth rates included in the sample according to the filtering applied by Cherp et al.<sup>30</sup> are shown as dots, each color corresponding to one country. The solid line shows the median maximum historical growth rate in the sample. Dashed lines represent the interquartile range of the distribution. This illustrates the non-stationarity of the distribution of  $G$  and that early forecasts based on the Cherp et al. method yield results that are too low by more than an order of magnitude.

Both of the tests for overfitting we have done so far relied on gathering new data. Based on the principles of backtesting, there are also opportunities to test for overfitting without adding new data. In Figure 3, the non-stationarity of the results is also visible by comparing predictions in 2019 with earlier measurements of the maximum growth rates. The red dots in Figure 3 correspond to the maximum growth rates measured in 2015. These are mostly below the identity line, demonstrating that the maximum growth rates still tend to grow with time, as one would expect for countries that have still not yet reached their true maximum growth rate.

Sensitivity testing could also have been used to make the underlying problem clear. This is demonstrated in Figure 5, where we compare fits for onshore wind power deployment in Austria, China, and Finland. We fit S-curves to the data from the first data point up to a maximum time  $t_{\max}$ , which we vary. As shown in Figures 5A–5C, for these countries, the resulting fits for  $k$  are wildly different from those for  $t_{\max} = 2019$ . As seen in Figures 5D and 5E, the fitted S-curve parameters also vary greatly. In Figure 5E, the estimates of the asymptotic deployment  $L$  reach levels as high as 3,500% of the national electricity supply. Similarly, the exponential growth rate  $k$  in Figure 5D shows large fluctuations between 120% and 20%. As a result, the estimated maximum growth rates ( $kL/4$ ) seen in Figure 5G reach levels as high as 400% per year.

Cherp et al.<sup>30</sup> make an effort to manage these instabilities by only considering countries that the S-curve fit suggests have already passed their point of maximum growth. This is done based on the requirement that  $m = x_{t_{\max}}/L > 0.5$ . However, as we see in Figures 5A–5C, this often results in countries going in and out of the selected sample of mature countries. Figure 5F shows the maturity  $m$  for Austria, China, and Finland for different years of evaluation  $t_{\max}$ , with the cutoff  $m = 0.5$  as



**Figure 5. S-curve fits for electricity generation from onshore wind in Austria, China, and Finland for different maximum years  $t_{\max}$**

The deployment of onshore wind in Austria (A), China (B), and Finland (C) shows intervals of rapid acceleration and deceleration. Consequently, the logistic S-curve fit based on data up until  $t_{\max} = 2012$  (green) differs starkly from that including data up until  $t_{\max} = 2019$  (orange). Both the exponential growth parameter  $k$  of the resulting S-curve fit (D) and the S-curve asymptote  $L$  (E) vary greatly depending on the length of the time series. As a result, the maturity  $m$  (F) and the maximum growth rate  $G$  (G) are also highly sensitive to the maximum year  $G$ . Figure S4 shows the equivalent figure using the Gompertz model, an alternative S-curve specification.

a reference. For their method to work as intended, one expects that once a country is deemed mature, it should generally remain in the sample, but this is not the case. Finland, for example, is classified as mature before 2010, then immature from 2011–2017, then mature from 2018 onward. Austria behaves similarly, leaving the sample in 2015 after many years of being included. China goes back and forth between included and excluded between 2007 and 2014.

In Figure S1, we show that we have not cherry-picked these examples. The estimated year  $t_0$ —where national growth should reach its maximum—increases systematically with time. On average, when  $t_{\max}$  increases by a year, the mean value of  $t_0$  increases by 0.6 years for wind and 0.3 years for solar. Across the length of the samples, the mean value of  $t_0$  increased by 8.4 years for wind and 2.1 years for

solar. This demonstrates that the answers yielded by their procedure are strongly non-stationary.

These problems stem from a combination of overfitting and selection bias. The overfitting occurs because there is not enough data and too much noise to fit a nonlinear function with three parameters. This means that the estimation of the parameters of the S-curve is not statistically significant, which, in turn, means that the determination of maturity based on the criterion  $m = x_{tmax}/L > 0.5$ , is not statistically significant either. The lack of statistical significance means that the selection of mature vs. immature countries is inconsistent. The fluctuations that indicate slow growth pass their test, whereas those that indicate rapid growth fail it, introducing selection bias. Only later, when the S-curves have reached maturity, do the fits become statistically significant, meaning that almost all countries pass the maturity test, and the sampling bias is removed.

### KEY TAKEAWAYS: MAKING RELIABLE STATISTICAL INFERENCES

We now summarize some of the key takeaways and suggest ways to avoid such problems in the future to produce more reliable technological change models.

The first takeaway is that statistical testing is essential. In all of the examples discussed here, statistical testing might have provided criteria for determining how reliable the models were. In the example of Cherp et al.,<sup>30</sup> assessing the statistical significance for their fits when judging whether a country has reached maturity could have made their conclusions more robust. However, there are two big challenges in doing so: the data for national diffusion are noisy, and the time series are short. There is a risk that under a meaningful test of statistical significance in measuring maturity, very few countries would have maturities with enough statistical significance to be useful at the present time.

This brings up another important lesson: some questions cannot be answered. There are many circumstances where there is insufficient data to answer questions quantitatively. In instances where statistical significance cannot be estimated, any answer needs to be treated with extreme caution. It is much better to say “we cannot reach a conclusion” rather than reaching a misleading conclusion.

An important caveat to statistical testing; however, is that it requires great care. Although many canned routines for fitting functions to data provide confidence intervals, these typically make strong assumptions, for example, the assumption that the data are independent and identically distributed (IID), which is often not appropriate for real data. For example, technology diffusion trajectories are usually strongly autocorrelated, meaning that the residuals from the S-curve fit are not independent, and the fluctuations tend to be very persistent. Moreover, S-curve growth is inherently non-stationary—the behavior during the early stages of a technology’s evolution is intrinsically different than the behavior during later stages. In part because of this, technology trajectories are heteroscedastic, meaning that the amplitude of the residuals varies with time, and consequently, the residuals are not identically distributed. These effects reduce the number of independent degrees of freedom and thus reduce statistical significance. As a result, methods assuming IID noise will often give over-optimistic answers that impart a misleading sense of confidence. Getting around this problem using in-sample tests is difficult and typically requires advanced numerical methods. (For an example of how statistical testing can be done when predicting technology costs, see Farmer and Lafond.<sup>34</sup>)



One of the most reliable ways to cope with this problem is to test conclusions out of sample. This is typically done by dividing the data into two or more statistically independent samples, performing model selection and parameter estimation on one (training) sample, and testing the resulting model on the other (testing) sample. This procedure has the huge advantage that it is not necessary to characterize exactly how the data deviate from being IID, which is usually difficult. It directly tests for overfitting, and it automatically takes deviations from IID data into account. Furthermore, it provides a straightforward way to measure how good the predictions are, which is valuable by itself: predictions without error estimates attached are not useful.

Although testing out of sample is the most reliable way to test for statistical significance, it still has pitfalls. While developing a model, researchers usually test many different models and select the one with the best performance. When tests are made repeatedly, this procedure becomes unreliable because the model's performance becomes positively biased, i.e., the model's score is inflated. The only test that is truly out of sample is the first one: as soon as a second test is made, the data are no longer out of sample—they have become part of the training set. To cope with this problem, it is good practice to divide the data into three samples for training, testing, and final validation. The training set is used to estimate the parameters of a model, and the test set is used for model selection. The final validation set is used only after a model has been selected and then should only be used once. If the final validation fails, then model selection must be repeated, with the corresponding danger of overfitting. Because of this, researchers should be very clear about their model selection protocol and ideally record the number of times that out-of-sample tests are carried out.

Data selection bias can emerge when data are selectively excluded, either from the model calibration process or in post-processing, i.e., after the model parameters have been fit. Reasons for excluding data are often justified, such as coping with obvious data errors or excluding outliers to make parameter estimation more robust. It can also occur implicitly when a dataset does not include certain data, for example, if data from developing countries are missing from a global study. The worst problems occur when selection causes systematic effects so that the data that are excluded have different properties than those that are not excluded. Such problems can be subtle, and researchers must think carefully about any procedures involving data selection.

Finally, researchers need to be more mindful of the bias-variance tradeoff, which refers to the fact that there are two fundamental reasons why models do not replicate reality exactly. A typical modeling exercise uses a functional form or "model family" (like the logistic S-curve) with free parameters (like  $k$ ,  $t_0$ , and  $L$ ) to fit the data (for this discussion, we use the term "model" to refer to a specific model with fixed parameters and "model family" to refer to a set of possible models determined by the free parameters, such as the logistic S-curve). Bias refers to the situation where no parameters of the model family exist that can match the real data. A common way to cope with bias is to introduce more parameters into the model family. However, this usually makes the errors in estimating the parameters worse, a problem that is called variance. Therefore, although there may exist parameter values that fit the data much better, it is impossible to find them because of larger estimation errors. This causes what is called the bias-variance tradeoff: the best model family is neither too simple nor too complicated; this happens when the errors caused by bias are roughly equal to the errors caused by variance.

Model users unfamiliar with this issue tend to overestimate the problem of bias and underestimate the problem of variance. When datasets are short and noisy, the best models are often much simpler than intuition might suppose. For example, for the problem of predicting technology costs, Nagy et al.<sup>28</sup> tested various hypotheses against data for 50 technologies; model families with two parameters made the best predictions, whereas all model families with three free parameters did worse.

To summarize, researchers working on technological change models must be careful to avoid overfitting and data selection bias by using statistical testing, ideally out of sample. It is essential to understand the reliability of predictions—this is usually a much harder problem than making predictions, but predictions of unknown quality are not helpful. Unsound inferences could lead to unsound policy advice, which risks unnecessary delays in the energy transition. Indeed, despite several predictions in the literature that we have highlighted here, there are no statistically sound S-curve forecasting methods that suggest that solar and wind energy cannot grow fast enough to meet the Paris Agreement targets or that this would be extraordinarily challenging in comparison with past technology growth trends. The procedures that we advocate above are routine in machine learning and for quantitative hedge funds, where survival depends on forecast reliability. The stakes for climate change are far higher, and the forecasting of climate mitigation deserves at least as much care.

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.joule.2024.07.016>.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## REFERENCES

- Grubb, M., Okereke, C., Arima, J., Bosetti, V., Chen, Y., Edmonds, J., Gupta, S., Köberle, A., Kverndokk, S., Malik, A., et al. (2022). Introduction and Framing. In IPCC 2022: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (Cambridge University Press).
- Krey, V. (2014). Global Energy-Climate Scenarios and Models: A Review. *WIREs Energy & Environment* 3, 363–383. <https://doi.org/10.1002/wene.98>.
- IPCC (2022). *Climate Change 2022—Mitigation of Climate Change—Full Report* (Cambridge University Press), pp. 1–30.
- UNFCCC (2015). *The Paris Agreement* (United Nations).
- Way, R., Ives, M.C., Mealy, P., and Farmer, J.D. (Sept. 2022). Empirically Grounded Technology Forecasts and the Energy Transition. *Joule* 6, 2057–2082. <https://doi.org/10.1016/j.joule.2022.08.009>.
- IRENA (2022). *Renewable Capacity Statistics 2022* (International Renewable Energy Agency).
- Grübler, A., and Nakićenović, N. (1991). Long Waves, Technology Diffusion, and Substitution. *Review* (Fernand Braudel Center) 14, 313–343.
- Farmer, J.D., Hepburn, C., Mealy, P., and Teytelboym, A. (2015). A Third Wave in the Economics of Climate Change. *Environ. Resource Econ.* 62, 329–357. <https://doi.org/10.1007/s10640-015-9965-2>.
- Geels, F.W., Sovacool, B.K., Schwanen, T., and Sorrell, S. (2017). The Socio-Technical Dynamics of Low-Carbon Transitions. *Joule* 1, 463–479. <https://doi.org/10.1016/j.joule.2017.09.018>.
- Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M.Z., Barrow, D.K., Ben Taieb, S., Bergmeir, C., Bessa, R.J., Bijak, J., Boylan, J.E., et al. (2022). Forecasting: Theory and Practice. *Int. J. Forecasting* 38, 705–871. <https://doi.org/10.1016/j.ijforecast.2021.11.001>.
- IEA (2023). *World Energy Balances*. <https://doi.org/10.1787/data-00512-en>.
- Ember (2024). *Yearly Electricity Data*. <https://ember-climate.org/data-catalogue/yearly-electricity-data/>.
- Hansen, J.P., Narbel, P.A., and Aksnes, D.L. (2017). Limits to Growth in the Renewable Energy Sector. *Renew. Sustain. Energy Rev.* 70, 769–774. <https://doi.org/10.1016/j.rser.2016.11.257>.
- Rypdal, K. (2018). Empirical growth models for the renewable energy sector. *Adv. Geosci.* 45, 35–44. <https://doi.org/10.5194/adgeo-45-35-2018>.
- IRENA (2024). *Renewable Capacity Statistics 2024*. <https://www.irena.org/Publications/2024/Jul/Renewable-energy-statistics-2024>.
- Madsen, D.N., and Hansen, J.P. (2019). Outlook of solar energy in Europe based on economic growth characteristics. *Renew. Sustain. Energy Rev.* 114, 109306. <https://doi.org/10.1016/j.rser.2019.109306>.
- Alavi-Koosha, A., Akbari, A., Toulabi, M., and Amraee, T. (2022). Trend Curve- and Machine Learning-Based Renewable Energy Development Forecast. In 2022 12th Smart Grid Conference, SGC, pp. 1–5. <https://doi.org/10.1109/SGC58052.2022.9998917>.
- Kramer, G.J., and Haigh, M. (2009). No Quick Switch to Low-Carbon Energy. *Nature* 462, 568–569. <https://doi.org/10.1038/462568a>.
- Höök, M., Li, J., Johansson, K., and Snowden, S. (2012). Growth Rates of Global Energy Systems and Future Outlooks. *Nat. Resour. Res.* 21, 23–41. <https://doi.org/10.1007/s11053-011-9162-0>.

20. Smil, V. (2016). Examining Energy Transitions: A Dozen Insights Based on Performance. *Energy Res. Soc. Sci.* 22, 194–197. <https://doi.org/10.1016/j.erss.2016.08.017>.
21. Iyer, G., Hultman, N., Eom, J., McJeon, H., Patel, P., and Clarke, L. (2015). Diffusion of Low-Carbon Technologies and the Feasibility of Long-Term Climate Targets. *Technol. Forecasting Soc. Change* 90, 103–118. <https://doi.org/10.1016/j.techfore.2013.08.025>.
22. Napp, T., Bernie, D., Thomas, R., Lowe, J., Hawkes, A., and Gambhir, A. (2017). Exploring the Feasibility of Low-Carbon Scenarios Using Historical Energy Transitions Analysis. *Energies* 10, 116. <https://doi.org/10.3390/en10010116>.
23. Hsieh, L., Pan, M.S., Chiang, Y.M., and Green, W.H. (2019). Learning only buys you so much: Practical limits on battery price reduction. *Appl. Energy* 239, 218–224. <https://doi.org/10.1016/j.apenergy.2019.01.138>.
24. Catsaros, O. (2023). Lithium-Ion Battery Pack Prices Hit Record Low of \$139/kWh. <https://about.bnef.com/blog/lithium-ion-battery-pack-prices-hit-record-low-of-139-kwh/>.
25. Gül, T., Pales, A.F., and Connelly, E. (2024). Global EV Outlook 2024: Moving towards increased affordability (International Energy Agency), p. 79. <https://iea.blob.core.windows.net/assets/a9e3544b-0b12-4e15-b407-65f5c8ce1b5f/GlobalEVOutlook2024.pdf>.
26. McKerracher, C. (2024). China's Batteries Are Now Cheap Enough to Power Huge Shifts. <https://www.bloomberg.com/asia>.
27. Penisa, X.N., Castro, M.T., Pascasio, J.D.A., Esparcia, E.A., Schmidt, O., and Ocon, J.D. (2020). Projecting the price of lithium-ion NMC battery packs using a multifactor learning curve model. *Energies* 13, 5276. <https://doi.org/10.3390/en13205276>.
28. Nagy, B., Farmer, J.D., Bui, Q.M., and Trancik, J.E. (2013). Statistical Basis for Predicting Technological Progress. *PLoS ONE* 8, 52669. <https://doi.org/10.1371/journal.pone.0052669>.
29. EIA (2023). Transportation Demand Module Assumptions. [https://www.eia.gov/outlooks/aeo/assumptions/pdf/TDM\\_Assumptions.pdf](https://www.eia.gov/outlooks/aeo/assumptions/pdf/TDM_Assumptions.pdf).
30. Cherp, A., Vinichenko, V., Tosun, J., Gordon, J.A., and Jewell, J. (2021). National Growth Dynamics of Wind and Solar Power Compared to the Growth Required for Global Climate Targets. *Nat. Energy* 6, 742–754. <https://doi.org/10.1038/s41560-021-00863-0>.
31. Nature Energy (2024). Article Metrics – National Growth Dynamics of Wind and Solar Power Compared to the Growth Required for Global Climate Targets. <https://www-nature-com.ezproxy-prd.bodleian.ox.ac.uk/articles/s41560-021-00863-0/metrics>.
32. Vinichenko, V., Cherp, A., and Jewell, J. (2021). Historical Precedents and Feasibility of Rapid Coal and Gas Decline Required for the 1.5°C Target. *One Earth* 4, 1477–1490. <https://doi.org/10.1016/j.oneear.2021.09.012>.
33. Nemet, G., Greene, J., Müller-Hansen, F., and Minx, J.C. (2023). Dataset on the Adoption of Historical Technologies Informs the Scale-up of Emerging Carbon Dioxide Removal Measures — Communications Earth & Environment. *Commun. Earth Environ.* 4, 397. <https://doi.org/10.1038/s43247-023-01056-1>.
34. Farmer, J.D., and Lafond, F. (2016). How Predictable Is Technological Progress? *Res. Policy* 45, 647–665. <https://doi.org/10.1016/j.respol.2015.11.001>.
35. Lafond, F., Bailey, A.G., Bakker, J.D., Rebois, D., Zadourian, R., McSharry, P., and Farmer, J.D. (2018). How Well Do Experience Curves Predict Technological Progress? A Method for Making Distributional Forecasts. *Technol. Forecasting Soc. Change* 128, 104–117. <https://doi.org/10.1016/j.techfore.2017.11.001>.
36. ITU (2022). ITU DataHub. <https://datahub.itu.int/data/?i=178>.