A Simple Computational Approach to Predict Long-Term Hourly Electric Consumption †

Eugene Pinsky *,‡, Etienne Meunier ‡, Pierre Moreau ‡ and Tanvi Sharma ‡

1. Introduction

With the emergence and proliferation of alternative energy sources, the ability to forecast long-term consumption is a growing key asset for companies and communities. It impacts strategic decisions such as energy pricing and investments in new power generation facilities. In the dynamic and diverse environment of smart grids, accurate long-term forecasting is an important practical problem.

While electricity consumption depends on factors like weather or income [1–3], we focus on long-term forecasting using only historical consumption. This paper focuses on predicting hourly energy consumption before the COVID-19 outbreak. The pandemic has drastically changed the patterns of energy consumption.

As historical data are often available—whereas matching weather and socio-economical data can be expensive, and adds complexity—using them to train our model is an asset in terms of scalability, generalization, and clarity. While short-term load forecasting is a mature field, long-term load forecasting algorithms are still too complex or not precise enough [4]. We are motivated to develop a simple algorithm that would allow people to understand the given long-term forecast better.

Our research expands beyond long-term consumption forecasting to include the sustainable integration of renewable energy sources. Using visual analysis, we assess the feasibility of meeting long-term consumption solely through renewable energy sources. This approach contributes to forecasting algorithm advancements and aligns with the transition towards sustainable energy solutions, offering valuable insights for strategic decision-making in energy planning and policy formulation.

1.1. Related Work

Electricity consumption forecasting is a highly active field in research. As described previously, the accurate forecasting of the consumption provides key advantages. Forecast-
ing is divided into multiple problems regarding the prediction length and data diversity. Short-term forecasting is focused on predicting the consumption during the next days, medium-term for the next weeks, and long-term for the next years. Information on an area’s weather conditions or economic activity is often used to improve the model but is not always available. Finally, some other techniques produce different possible forecasts based on scenarios. Each forecast is then linked to events that can happen in the future.

Currently, short-term forecasting \cite{5,6} represents most of the papers written. This can be explained by the fact that the forecasting of an hour or a day ahead of production will generally be more accurate and will still give a tactical advantage to the organization. However, several researchers have developed techniques to improve long-term forecasting (see “long-term electricity forecast: a systematic review” \cite{7}). Successful techniques often take into account other factors like the weather or the economic situation of the country, for example, in (Daneshi et al.) \cite{8} or (Goude et al.) \cite{9}.

Factor-based and time-series models offer distinct approaches for predicting energy consumption, each with their advantages and drawbacks. Factor-based models provide a comprehensive analysis by incorporating various influencing factors, offering insights into causality and adaptability but suffer from complexity and high data demands, raising the risk of overfitting. On the other hand, time-series models focus on historical consumption patterns to forecast future demand, boasting simplicity and effectiveness for short-term predictions with fewer data requirements. However, they may struggle with accounting for sudden changes due to external factors and capturing long-term trends. The choice between these methodologies depends on the specific forecasting context, data availability, and the forecast’s timeframe, with a hybrid approach sometimes serving as the most robust solution.

These methods show good results but imply the knowledge of numerous other factors, which are often hard to predict. By contrast, in our work, we are trying to forecast long term, based solely on the loads from previous years. We emphasize simple computational approaches that are easy to implement and explain.

Auto-regressive models are often used for this kind of task; for example Khuntia et al. \cite{10} use a Multiplicative Error Model and an estimation of the volatility to forecast the long-term electricity consumption with a monthly granularity. They manage to have a 7.04% MAPE on a 3-year out-of-sample forecasting using the PJM dataset \cite{11}. Safdarian et al. \cite{12} use a three-stage hierarchical approach (yearly, weekly, and daily trend models) and then the Box-Jenkins method to find the weekly and daily trend. This method can depict the pattern and accurately forecast the Finland average daily dataset. Newer techniques have also been applied to long-term load forecasting. The use of recurrent neural networks \cite{4} and feature engineering, breaking each date into hour, weekday, month, and year, managed to have a MAPE of 6.54% on a 4-year out-of-sample prediction with an hourly granularity on the ISO New England electricity market dataset \cite{13}. However, using recurrent networks for this task can create an important computational overhead that can be problematic if we want to make numerous predictions.

Launched in 2012, the Global Energy Forecasting Competition \cite{14} is aimed to compare different techniques for forecasting energy consumption. The qualifying phase of the competition, where it was requested to perform an hourly forecasting of the consumption for one month, generated several publications. In the publication from Ziel \cite{15}, the author split the load data into a long-term trend component and a remainder term, then used a moving average type model to model the trend component and a quantile regression model for the remainder term with one different model for each hour (24 different models in total), and the result was evaluated using the pinball score. Some other papers focused on machine learning techniques to produce forecasting for this task; for example, ref. \cite{16} used an ensemble with tree-based algorithms and neural networks to achieve good results.
1.2. Quick Summary

The method presented in this paper is designed for long-term forecasting and incorporates the seasonality of the historical load consumption. Our computational procedure is simple and amenable to an intuitive interpretation. Moreover, it can be easily generalized for forecasting at different time scales.

To convey a challenge in predicting hourly consumption, consider the difference in such consumption between a weekday (e.g., Wednesday) and the weekend (e.g., Saturday). Figure 1 illustrates the hourly breakdown of electricity generation during specific periods—first Saturdays throughout 2016 for weekends and second Wednesdays in 2016 for weekdays. The data presented suggest a marked difference between weekend and weekday electricity production. Generally, electricity consumption tends to be lower on weekends and holidays due to commercial offices’ closure and reduced demand for lighting and computer equipment. In the electricity industry, usage periods are typically categorized as on-peak hours, spanning from 7:00 a.m. to 11:00 p.m. on weekdays, and off-peak hours, covering 11:00 p.m. to 7:00 a.m. on weekdays, as well as the entire day on Saturdays, Sundays, and holidays.

![Figure 1. Weekday vs. weekend electricity consumption.](image)

Let us present an intuitive description of our algorithm (its pipeline is presented in Figure 2 and will be discussed in detail in Section 3). If we examine the RTE dataset (consumption in France from 2008–2011) in Figure 3, we can see that the power consumption is periodic each year. At the same time, total power consumption from year to year could change, reflecting new technologies’ emergence. Our approach consists of two steps:

1. Remove the major trend reflecting changes in the annual consumption. We do this by using a simple linear autoregression (we also tried a second-degree order linear auto-regression for some of the datasets). Note that after the trend is removed, it is still possible that the detrended time series is seasonal, and hence, non-stationary.

2. We predict daily consumption for each day by looking at the appropriate average of the power consumption for “similar” days in the past.

For example, we want to predict the 3 April 2020 power consumption. This date is the first Friday in April. We look at historical data and find all days that are the first Fridays in April. These are our “similar” days. We take these “similar” days and compute the average of their consumption for previous years. Intuitively, power consumption has two components: (1) major trend, and (2) daily consumption that depends strongly on seasons, months, and days of the week. By averaging over “similar” days, we can capture this periodicity simply and intuitively.

From (1) and (2), we compute our prediction by adding the global trend for that day and the average of “similar” days. Note that we can easily extend the proposed procedure for any time scale, such as hours, by averaging over the “similar” hours in the historical data. Moreover, the procedure can be implemented directly in SQL.
2. Mathematical Notation and Preliminaries

2.1. Detrending

Removing the yearly trend allows the time series to be closer to stationarity. With this in mind, we can decompose the consumption into trend, seasonality, and error. The first step is to remove the annual trend. Removing the yearly trend is a step towards enhancing the stationarity of the time series, although the presence of seasonality can still result in a non-stationary series.

A simple approach is to consider a linear auto-regression. First-order linear regression is adapted if the consumption is constantly increasing or decreasing yearly. However, in the event of an economic crisis, for example, we may need to use a polynomial or other methods to model the trend. Other detrending methods, like the join-point model, are too computationally expensive for our use cases. Therefore, we will use first- or second-order linear auto-regression.
Once we compute the trend, we can subtract it from the consumption to have a yearly stationary time series. We can then perform two different forecasts: for the seasonal part, and for the global trend. Once these forecasts are computed, we take their sum and use it as our final forecast.

2.2. Mean Absolute Percentage Error

In order to quantitatively evaluate our forecasts, we use the MAPE as our performance metric. We choose this metric because it is easy to understand, gives ideas on how accurate the forecast is, and is commonly used for load forecasting problems. It allows us to compare our results with the ones given by other methods:

\[
MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right|
\]

where \( n \) is the number of values to forecast, \( A_i \) is the actual value, and \( F_i \) is the forecast value.

2.3. Confidence Interval

The use of the confidence interval gives us information about the regularity of errors and can bring more confidence into the forecaster. To calculate the confidence band (CB), we use:

\[
CB = [\bar{X} - t_{\alpha} \frac{Std(X)}{\sqrt{n}}; \bar{X} + t_{\alpha} \frac{Std(X)}{\sqrt{n}}]
\]

where \( \bar{X} \) is the vector of the mean absolute error, and \( n \) is the number of samples. We consider \( t_{\alpha} = 1.96 \) since we want the 95% confidence interval and have more than 100 samples.

The construction of the CB presupposes that the mean absolute errors are independent. This assumption allows us to apply the central limit theorem, justifying the use of the normal distribution to construct the confidence interval around the mean of the MAEs. This approach assumes that, with a sufficiently large sample size, the distribution of the mean absolute errors will approximate a normal distribution, enabling the application of the t-distribution for the confidence band calculation. We acknowledge that these assumptions are idealized and that the actual distribution of errors may deviate from normality, especially for smaller sample sizes or non-linear error distributions. However, given that our sample size exceeds 100, the approximation is considered reasonable for our analysis.

3. Development

3.1. Datasets

Three different datasets were used:

- The French Réseau de Transport d’Électricité (RTE) company, in charge of the French power distribution network, published its dataset of the overall half-hour consumption across all France from 2008 to 2018 [17].
- The same French RTE company published its dataset of the overall production of electricity produced by renewable resources in France (except Corsica) from 2012 to 2020.
- The American Independent System Operators (ISO) company, operating New England’s grid and in charge of power system planning, used to publish a dataset of hourly consumption in New England from 2004 to 2015.
- The American Pennsylvania-New Jersey-Maryland Interconnection (PJM) organization [11] coordinates the movement of wholesale electricity in all parts of 13 US states. The organization published datasets of metered data aggregated from the zones’ respective electric distribution companies. We can find the datasets of hourly zone loads from as early as 1993, depending on the zones.
3.2. Pre-Processing

During pre-processing, we remove all unwanted columns from the datasets and keep only the date time column and the \texttt{power\_consumption} column. We then re-sample those datasets on hourly measures, and fill the missing values with a forward fill. The forward fill is a procedure that fills the missing data using the last known measure. We choose this method after verifying that short-term naïve predictions (i.e., 1 h forward propagation) are precise enough (e.g., overall MAPE of 3.7% on the RTE dataset) and that we do not have multiple missing values in a row in our datasets.

3.3. Schema of the Pipeline

The Python source code for the pipeline is available on our GitHub [18]. Its schema is given in Figure 2.

3.4. Detrending Explanation

By removing the trend that characterizes the evolution of power consumption between the different years, we can focus on the seasonality. To perform that, we run a linear regression using Mean Square Error.

The detrending technique shown in Figure 3 extracts the linear trend which is enough for most of the datasets. In the PJM dataset, where we have an augmentation until 2008 then a change in trend due to the economic crisis, we do not have a linear trend. We can use a polynomial regression to model the trend. This is shown in Figures 4 and 5.

![Figure 4. Detrending PJM dataset using 1-degree polynomial.](image)

![Figure 5. Detrending PJM dataset using 2-degree polynomial.](image)
3.5. Model Year Explanation

3.5.1. Data Analysis

Power prediction is a domain where data are highly seasonal and patterns repeat themselves. For instance, if we plot the power consumption in regards to the month or the weekday, the curves are all within a narrow confidence interval, cf. Figure 6.

![Figure 6. Detrended RTE electricity consumption by months.](image)

Those results are easy to explain, as overall consumption is lower on non-working days, and one of the main power consumption factors is heating, without or close to no Air Conditioner (AC) consumption at all in the summer. On the ISO dataset for instance, we see the same correlations even though there is a consumption peak in summer due to the heavy AC usage.

However, the same plots based on the day number in Figure 7 do not give any insight, so the number of the day in the month has no direct impact on the electricity consumption.

More generally, we can use the Autocorrelation Function (ACF) to extract all those autocorrelation lags for us.

![Figure 7. Detrended RTE electricity consumption by day numbers.](image)

The correlograms in Figure 8 show us that there are strong autocorrelations in the time series with lags every 24 h, 7 days, and 365 days. We can leverage all the autocorrelation lags we found by grouping the historical data on those attributes and using the average of each group to predict the consumption.

![Figure 8. ACF tests on the RTE dataset at different lags.](image)
3.5.2. Building the Model

Figure 9 shows a simplified representation of our model year approach. The top quadrants plot the historical consumption data grouped by months $m$ and weekdays $w$, and show the distinction over years (here only for January and June).

![Figure 9. Simplified representation of the model year computation with the RTE dataset.](image)

If we take the average of the historical data grouped by those attributes ($m$ and $w$), this gives us the orange dotted prediction line on the bottom quadrants. We see that the resulting prediction is quite close to the real values in June but quite far in January.

With the same logic, we can improve this prediction by adding more levels of detail, using the granularity attributes identified during the correlation analysis above.

3.5.3. Formula

If we want to predict the consumption on Sunday, 2 April 2023, then we take the average consumption for every Sunday of the first week of April from all historical detrended data. We then add the projection of the global trend at that day.

In detail, considering $Y = \{ y_t, t = 1, \ldots, T \}$ the set of past recorded $T$ years, the predicted power consumption $\hat{X}$ for year $y_{T+k} \notin Y$, month $m$, part of month $n$, weekday $w$, and hour $h$ based on the recorded power values $X$ and the global trend $G$ is:

$$\hat{X}_{y_{T+k}, m, n, w, h} = \frac{1}{T} \sum_{t=1}^{T} X_{y_t, m, n, w, h} + G_{y_{T+k}, m, n, w, h}$$  \hspace{1cm} (1)

4. Results

Using the model year we created during the training, we can now have a forecast of the load consumption for the next years.

As we will see, our approach is able to obtain the different patterns that compose the load consumption through the year. Using the trend forecasting and retrending step, we are able to have a long-term forecast very close to the real consumption.

4.1. RTE Dataset

With the RTE dataset, using only 4 years of data as the training set (2008–2011), we compare our forecast models against the next 6 years of consumption (2012–2018). The results are shown in Figure 10 and Table 1.

A limitation of our method is that aside from the global trend, all our future years have the same forecast, as they are all predicted using the same year model. This limitation is clearly visible, for example, in Figure 10. However, the errors are still good considering we are forecasting the next 6 years of hourly consumption based solely on the last 4 years.
Figure 10. Daily RTE consumption (top) vs. forecast (bottom) for years 2013–2014, using years 2008–2011 as the training set.

Table 1. RTE hourly predictions metrics using years 2008–2011 as the training set.

<table>
<thead>
<tr>
<th>Year</th>
<th>Our MAPE</th>
<th>Our Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>6.79</td>
<td>10.07</td>
</tr>
<tr>
<td>2013</td>
<td>5.75</td>
<td>8.95</td>
</tr>
<tr>
<td>2014</td>
<td>5.1</td>
<td>8.09</td>
</tr>
<tr>
<td>2015</td>
<td>5.38</td>
<td>8.11</td>
</tr>
<tr>
<td>2016</td>
<td>5.51</td>
<td>7.77</td>
</tr>
<tr>
<td>2017</td>
<td>5.88</td>
<td>8.47</td>
</tr>
<tr>
<td>2018</td>
<td>6.23</td>
<td>12.89</td>
</tr>
<tr>
<td>Overall</td>
<td>7.07</td>
<td>4.71</td>
</tr>
</tbody>
</table>

4.2. ISO Dataset

We also perform an out-of-sample forecast on the ISO dataset for the 5 years of 2010–2015, with a training set composed of the data from the 5 years 2004–2009. Table 2 shows our results compared to the results made by a LSTM with the same train and test sets and a very similar feature extraction process in “Long term load forecasting with hourly predictions based on long-short-term-memory networks” [4].

Table 2. ISO hourly predictions metrics vs. results from [4] using years 2004–2009 as the training set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>5.69</td>
<td>8.31</td>
<td>5.6</td>
<td>4.8</td>
</tr>
<tr>
<td>2011</td>
<td>5.62</td>
<td>8.02</td>
<td>7.5</td>
<td>1.61</td>
</tr>
<tr>
<td>2012</td>
<td>6.18</td>
<td>8.67</td>
<td>6.6</td>
<td>6.35</td>
</tr>
<tr>
<td>2013</td>
<td>6.11</td>
<td>8.29</td>
<td>6.6</td>
<td>5.19</td>
</tr>
<tr>
<td>2014</td>
<td>6.5</td>
<td>7.85</td>
<td>6.17</td>
<td>6.38</td>
</tr>
<tr>
<td>2015</td>
<td>7.21</td>
<td>11.44</td>
<td>6.54</td>
<td>2.25</td>
</tr>
<tr>
<td>Overall</td>
<td><strong>6.34</strong></td>
<td><strong>4.99</strong></td>
<td><strong>6.54</strong></td>
<td><strong>2.25</strong></td>
</tr>
</tbody>
</table>
The wider confidence intervals from our method compared to another may be due to our model accounting for a greater range of uncertainties and being more sensitive to data fluctuations. This conservative approach reflects a more cautious estimate of forecast variability. Although wider intervals suggest higher uncertainty, they do not inherently indicate lower forecast accuracy but rather a more cautious representation of the forecast’s potential variability. Further model refinement could help narrow these intervals while maintaining a realistic assessment of uncertainty.

As we can see here, our results are comparable with the LSTM. The advantage of our approach is that it is intuitive and fast to compute. In this example from paper [4], “the model was trained using a 12 GB NVIDIA Graphics Processing Unit (GPU) and required a computation time of approximately 30 min”, whereas it only took our method a few seconds on a laptop without using a GPU.

Zooming on the hourly forecasting for January 2013 (i.e., 4 years after the last training data) gives Figure 11, where we can observe the hourly, daily, and weekly pattern.

![Figure 11. Hourly ISO consumption (top) vs. forecast (bottom) for January 2013, using years 2004–2009 as the training set.](image)

4.3. PJM Dataset

PJM offers us the hourly load consumption of many load zones across the USA. This dataset spans from 1993 to 2018, which is very interesting because this covers the great economic recession in 2008 that impacted the load consumption by changing its global trend; this shows an example where the model cannot predict accurately using a linear trend.

We forecast the 5 years of 2010–2014 from the 18 years of 1993–2010 in the training set. We can see the results in Figure 12 and Table 3.

Table 3. PJM hourly predictions metrics using years 1993–2010 as the training set with a 1-degree linear detrending.

<table>
<thead>
<tr>
<th>Year</th>
<th>Our MAPE</th>
<th>Our Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>9.73</td>
<td>9.33</td>
</tr>
<tr>
<td>2011</td>
<td>11.07</td>
<td>12.27</td>
</tr>
<tr>
<td>2012</td>
<td>11.68</td>
<td>12.31</td>
</tr>
<tr>
<td>2013</td>
<td>10.96</td>
<td>10.59</td>
</tr>
<tr>
<td>2014</td>
<td>10.38</td>
<td>14.11</td>
</tr>
<tr>
<td>Overall</td>
<td>12.67</td>
<td>7.88</td>
</tr>
</tbody>
</table>
Improving the Detrending

This is not very good because the 1-degree polynomial we use to detrend the time series does not model well the yearly trend, which is increasing before 2008 and then decreasing after the economic crisis (cf. Figure 4).

Table 4 shows our results for the same forecasting but using a 2-degree polynomial for detrending instead (cf. Figure 5). As we can see, the results are already considerably better because the detrending function is more adapted to the yearly trend of the time series.

Table 4. PJM hourly predictions metrics using years 1993–2010 as the training set with a 2-degree polynomial detrending.

<table>
<thead>
<tr>
<th>Year</th>
<th>Our MAPE</th>
<th>Our Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>8.28</td>
<td>8.72</td>
</tr>
<tr>
<td>2011</td>
<td>9.35</td>
<td>11.39</td>
</tr>
<tr>
<td>2012</td>
<td>10.00</td>
<td>11.45</td>
</tr>
<tr>
<td>2013</td>
<td>9.38</td>
<td>9.49</td>
</tr>
<tr>
<td>2014</td>
<td>9.05</td>
<td>9.49</td>
</tr>
<tr>
<td>Overall</td>
<td>10.04</td>
<td>6.80</td>
</tr>
</tbody>
</table>

We can then compare our results with the one from a study using a Multiplicative Error Model to perform long-term forecasting on this dataset with a monthly granularity. They managed to have an error of 7.09%, while we have an error of 10.07% using a 1-degree polynomial for detrending and 7.69% using a 2-degree polynomial for detrending.

5. Renewable Energy Sources

We conclude our paper with a short discussion on renewable energy. The French energy landscape has undergone a remarkable transformation in recent years, with a slight but notable surge in the adoption of renewable energy sources. From 2012 to 2020, France experienced a noteworthy growth trajectory, positioning itself as a leading advocate for clean and sustainable energy practices to fulfill its commitment to become a country with “net zero carbon dioxide emission” by 2035.
Traditionally known for its reliance on nuclear power, France has embraced a paradigm shift towards renewable energy sources, demonstrating a commitment to a more sustainable and environmentally friendly energy landscape. Figure 13 shows the distribution of electricity production by various sources. The solar revolution and the wind expansion can be seen in this figure. France’s approach to renewable energy involves a multi-faceted strategy encompassing solar, wind, hydropower, and biomass.

![Figure 13. Distribution of total energy production by source.](image)

Figure 14 shows the breakdown of total energy production to renewable and non-renewable from 2012 to 2019. The total contribution of renewable energy sources increased from 16.4% to 21.6% over this period. Non-renewable sources, such as nuclear, gas, etc., dropped their contribution from 83.6% to 78.4%, a difference of 5.2%, in this period. The majority of the increase in renewable energy is by solar, followed by wind and biomass production.

![Figure 14. Overall distribution of energy production.](image)

Renewable energy is still a minor component in total energy production, but it is growing steadily. We plan to focus on forecasting renewable energy in our subsequent work.

We understand some of the limitations of the proposed approach. First, we analyzed only France. This may not be a good example of renewable energy because most of their electricity is generated by nuclear power. Secondly, it is difficult to estimate how our method (or competing methods) performs under shocks such as the COVID-19 pandemic or wars. We hope to address these issues in our future work.

6. Conclusions

The seasonality of the load forecasting domain allowed us to make a good prediction model that scales well. Our results are competitive with some heavier machine learning algorithms like regressors or LSTMs. Our approach is focused on simplicity and scalability, without the need for complex third-party code libraries. The prediction itself made by the model is intuitive and easier to understand, debug, and interpret. Getting a grasp of
how the prediction is computed based on the identified correlated attributes gives the user intuition about how electricity consumption is evolving.

Our model year algorithm is simple enough to be implemented in SQL and makes accurate long-term predictions based on huge historical datasets in online analytical data warehouses. We believe that such an intuitive and simple approach can be applied to many time-series forecasting models.

We focused on a dataset before the outbreak of the COVID-19 pandemic, which radically changed the patterns of energy consumption for a few years. In the aftermath of the pandemic, we expect the suggested models to offer simple, intuitive, and accurate predictions.

Author Contributions: All authors contributed equally to this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.


Acknowledgments: The authors would like to thank the Metropolitan College of Boston University for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.