Master's in Industrial Technologies

Master's Final Thesis

# Solar and Wind Power Generation Modeling for the Spanish Electrical System

Author: Fernando Celaya

Supervisors: Andrés Ramos and Jesús David Gómez

October 2, 2024

**Abstract**

INTRODUCE ABSTRACT

***Keywords***— INTRODUCE KEYWORDS

# Acknowledgements

WRITE ACKNOWLEDGEMENTS

# Contents

# List of Figures



# List of Tables

# 1 Introduction

## 1.1 Background

For the last few decades, the global energy landscape has been undergoing a significant transformation, driven by climate change and the need to reduce greenhouse ges emissions in order to reduce its impact. Europe has been at the forefront of this transformation and Spain, as its fifth largest economy, has had a significant part in this. In fact, Spain has shown drive of its own by being at the forefront of many of these initiatives with a robust commitment to renewable energy sources. This country has set ambitious targets for renewable energy integration, seting a target of having 74% of its energy coming from renewable generation facilities by 2030 and a 42% share of renewables in energy end use, as per the 2021 Spain Energy Policy Review Agency, 2021. Spain is particularly well positioned for this transition due in part to its favourable geographical conditions. Being one of the southermost parts of Europe with a mediterranean climate allows for long periods of intense sunshine, and its extensive coast also provides good conditions for regular and powerful wind.

Like any other industrialised country, historically Spain has relied heavily on fossil fuels. The recent growth of renewable energy has been driven by several factors, including technological advancements, a reduction in costs driven by the economies of scale in production mainly in China and by government policies. Amongst these policies, the first big push came from the Royal Decree 661/2007 through which the production of renewable energy in Spain was regulated T. y. C. Ministerio de Industria, 2007. In this legislation, a premium was awarded to producers of renewable energy in order to incentivize investment in the sector. However, as soon as 2010 these premiums started to be reduced due to the general deficit in the sector and the financial constraints imposed on the state by the 2008 financial crisis. These reductions culminated in 2013 with a drop in 40% of all premiums available at that point. This led to the bankruptcy of a great number of companies which had invested heavily mainly in solar power plants and which relied on these subsidies for their required profitability. In 2015 the sector was further regulated through what came to be known as the "Sun Tax" E. y. T. Ministerio de Industria, 2015. This royal decree imposed a tax on self-consumption systems – individuals or companies installing solar PV panels to generate part of their own electricity – in order to compensate for the additional system maintenance needed by these systems. Since then, thanks to the drop in costs these subsidies have stopped being necessary with solar energy being profitable without government intervention, and some argue even the most profitable source of energy Glenk and Reichelstein, 2022. This advantageous financial landscape, together with a recent regulatory push driven mainly from the European Union, has prompted a new golden age of investment in renewable energy in Spain. In Figure 1 the growing trend in investment can be clearly seen, except for the drop in 2021 created by the Covid-19 crisis.

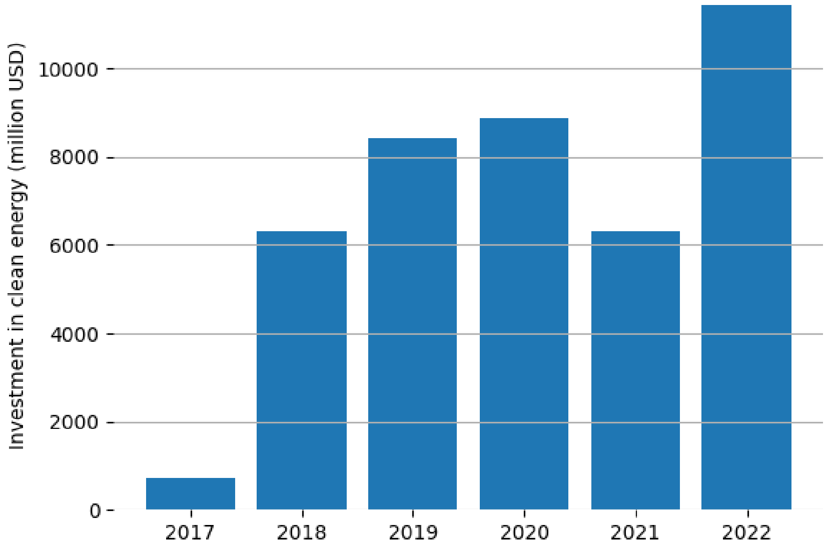


Figure 1: Investment in the Spanish clean energy sector, in million USD. Fernández, 2024

This showcases the obvious interest there is in renewable energy in general in Spain. However, when talking about renewable energy it is very important to outline the different renewable energy resources and their corresponding technolgies Ellabban et al., 2014:

- **Biomass**: Biomass energy is derived from organic materials, mainly plant and animal matter such as wood, agricultural crops, waste, and algae. It can be burned directly to produce heat and power a thermal power plant or it can be converted into biofuels like bioethanol or biodiesel. These fuels are then used in thermal power plants or for other uses like vehicle powering. It is very versatile with many of the main benefits of fossil fuels and can be used to reduce waste. However, sustainable sourcing of the biomass is critial, as otherwise it can lead to deforestation or other environmental issues.

- **Geothermal**: Geothermal energy harnesses the higher and more constant temperatures found deep below the Earth's surface, which originates from the planet's hot core due to the high pressure and the radioactive decay of different materials. Geothermal plants directly source hot water from underwater sources or pump fluids deep into the surface, heating it and retrieving it afterwards. This hot fluids can be used in a thermal power plant, as part of the cycle to preheat the working fluid or directly to heat up some buildings. It provides a constant and reliable energy supply not dependent on weather conditions and it has a samll land footprint. However, it is very location-specific, working best in regions with significant tectonic activity and with easily drillable materials in the ground.

- **Hydropower**: Hydroelectric power is the power of flowing water. This water is generally stored in a dam and released to flow through turbines which generate electricity, although other systems which do not require dams and use water flowing in rivers are also possible. Worth mentioning are also pumped storage plants, which although are not strictly a renewable energy resource can be used as energy storage systems. Hydroelectric power has high efficiency and can provide electricty on demand with very short warm up periods, however the initial investment to build the dam is very costly, it cannot be placed anywhere and can disrupt the local ecosystems.
- **Marine**: Marine power leverages different sources of energy available in the seas and oceans. There are six main energy sources within this category:
  - Wave: It transforms the kinetic energy of waves in their up and down movements into electrical energy.
  - Tidal range: It leverages the difference in height between the high and low tide to generate electricity.
  - Tidal currents: It harnesses the horizontal currents caused by the rising and falling of the tide.
  - Ocean currents: Similar to the previous source, however it leveraes currents not necessarily caused by tides, but caused by the dynamics of the ocean.
  - OTEC: Ocean thermal energy conversion exploits the temperature difference between the warmer surface water and the colder deep water as the hot and cold sources of a thermal cycle to generate electrictiy.

  - Salinity gradients: It leverages the difference in saline concentration between different areas of the ocean.

  Many of these energy sources are still undergoing intensive research and developement and are still in the prototype and demonstration stage.

- **Solar**: Solar energy leverages the energy of solar radiation reaching us from the sun in several different ways:

  - Photovoltaic: Solar PV systems directly convert the energy of the photons sent by the sun into direct current through the photoelectric effect Einstein, 1905. The PV cells, generally made out of silicon, have an efficiency of around 20%. However they are very modular and can be produced at scale, helping achieve the tremendous drop in cost that has been seen during the last decade.
  - Thermal: Solar thermal power leverages the heating capacity of the solar radiation to harness its power. Although it is often used to directly heat up buildings, it can also be used to generate electricity. Concentrating solar power (CSP) produces electricity by concentrating the solar irradiance through mirrors and lenses to heat a liquid which is used in a thermal power cycle.

- **Wind**: Wind energy refers to the kinetic energy of wind currents. This energy can be transformed through wind turbines into electrical energy. Wind power is generally obtained through two different types of setups. There are on-shore and off-shore power plants, with the former being installed in land and the latter in seas or oceans near the coast. Off-shore installations have some advantages, like more powerful and constant wind, no land erosion and less landscape visual degradation, however its installation is more costly.

As it can be seen, there are many different renewable energy resources and technologies that can be used to harness them and aid in the energy transition. However, not all of them are equivalent, with different technologies having different levels of maturity, profitability, capital requirements, etc. That is why generally the renewable energy mix is not very uniform.

In Figure 2 it can be seen how solar, wind and hydro energy covers more than 95% of the energy generation of 2023 – as percentage of the total generation of 134.321 GWh of renewable

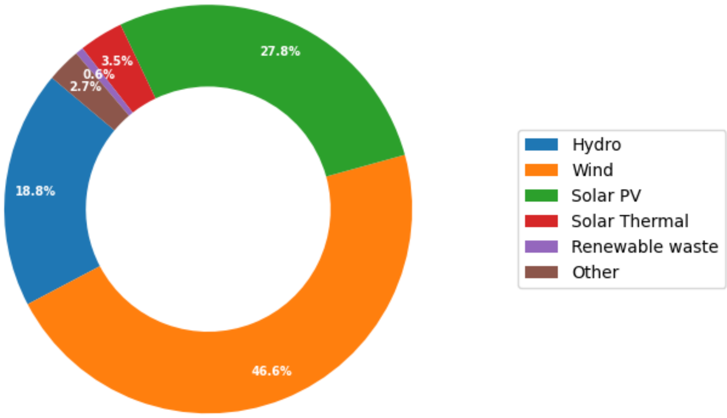


Figure 2: Generated renewable electricity mix in Spain in 2023. *Renewable electricity generation*, 2024

energy – while other sources like geothermal and biomass are almost negligible.

## 1.2 Motivation

In the previous section in Figure 2 a promising picture of the renewable energy generation in Spain could be seen, even more when the total generation is compared with the non-renewable generation of 132.486 GWh in Spain *Renewable electricity generation*, 2024, showing how renewable energy generation surpassed the 50% mark.

However, the generation mix shown is also a source for concern. A 77.9% of the total energy mix was produced through solar and wind energy. Solar and wind energy share some characteristics which lead to some grave challenges to the overall electrical system. Mainly, these characteristics and consequent challenges are:

- **Variability**: These energy sources are intermittent. That is, one cannot choose to turn on a solar power plant at any moment, as its output depends on the solar input, which cannot be decided by the operator. This is translated into a very variable pattern of generation with high peaks and low troughs.
  This variable pattern of production with rapid ramps leads to the need of having other power plants with also rapid ramp ups which can step up after a sudden fall of solar or wind production. These other plants also need to be very flexible to adapt to the wide changes in generation from these renewable resources.

- **Uncertainty**: Not only can these technologies not be discretionarily dispatched, but often it is difficult to know in advance what their behaviour will be at some point in the future. The further out into the future one looks, the less certain the sun and wind forecasts become, and thus the less certain the production output forecasts become.

Many of the problems of variability are accentuated by the fact that this variability is uncertain. Given the grid players are uncertain about the renewable generation, large reserves need to be allocated in case the prediction fails. This in turn leads to more expensive electricity for the end consumer if the same level of reliability is required.

- **Location specific**: The renewable resources are often geographically concentrated. For example, sun is more prominient in the south of Spain and wind around the coast.

  Having many of these power plants geographically concentrated leads to several problems when all of them are producing at high outputs at the same time. Power lines connecting the renewable rich area with the rest can become overloaded forcing shutdowns and cutoffs. Furthermore, great voltage differences between areas in the grid can appear.

- **Generator technology**: The induction generators generally used in wind turbines or the inverters in solar power plants behave differently to the synchronous generators used in thermal power plants, which leads to problems in frequency control as will be outlined below.

  This difference in generator technology leads to several problem. The first one being that these power plants are unable of providing the primary reserve matching supply and demand at the expense of frequency changes that synchronous generators supply automatically. Furthermore, these generators do not provide the same reactive power supply as the synchronous generators, and in the case of induction generators in wind turbines they in fact consume reactive power in order to function. It has also been hypothesized how these generators can lead to angular instability Vittal and Ayyanar, 2013. The generators of wind turbines have also been shown to lead to problems of power quality due to the injection of different harmonics Muljadi et al., 2006.

- **Low capacity factor**: Due to the variability in the resources, the capacity factor of installations leveraging sun and wind is generally lower to that of other non renewable technologies.

Even more importantly, the greater the share of overall energy that comes from solar and wind sources the more pronounced these challenges will become. For more details regarding these characteristics and their consequences, refer to Ahmed et al., 2020, Kumar et al., 2016 and Steen et al., 2014.

Grid operators need to consider these challenges at a short and medium term when opperating the grid, and plannificators should also consider it when planning the long term of the grid.

The first two characteristics, the variability and uncertainty of these resources, leads in fact to a great problem in long term planning. When deciding where to locate which power plants and of what capacity, it is important to have accurate models of all resources. In fact, many of the optimization models used in long term planning require forecast models of these resources, which is very challenging due to the aforementioned characteristics. The lack of realistic and accurate models of solar and wind energy generation is precisely what this work aims to solve.

## 1.3 Objective and scope

By looking at the overview of the wind and solar energy presented in subsection 1.2 Motivation and understanding the objectives and challenges of grid planners – long term distribution of resources to ensure the grid's effectiveness and reliability – the need for a model that can be used by these planners has become apparent.

To be more precise, for such a model to be useful it must accurately represent the wind and solar energy generation marginal and joint distributions. The characteristics the model must fulfill are the following:

- **Modeled variable**: The variable to be modeled is the hourly capacity factor of wind, solar PV and solar thermal energy generation. The capacity factor $f_t$ can be calcualted as

$$f_t = \frac{E_t}{P_t t} \tag{1}$$

  Where $E_t$ is the energy generated in a given timeframe, $P_t$ is the power output rating – that is the rated installed capacity – and $t$ is the length of the timeframe. As it can be seen, the capacity factor is a unitless ratio.

  It has been chosen as the modeled variable due to having several beneficial characteristics. Numerically, $f_t \in [0, 1]$ which makes it easier to model. It also is able to show the power generation without regarding the installed capacity. The installed capacity is often determined by policy decisions, awarded licenses or discretionary investment decisions by electricity companies and is thus harder to accurately model. The capacity factor however allows for an accurate representation of the generated power.

The hourly frequency has been chosen as it is short enough to allow for significant changes throughout the day, but is not short enough for noise to become a significant part of the variable. Furthermore, this is the frequency used by most market players, with the electrical market quoting hourly values for example.

The wind, solar PV and solar thermal capacity factors have been chosen as independent variables because the three of them have significantly different driving factors, making them have different dynamics that can be captured by individual models. Even the solar PV and solar thermal have different dynamics, thanks to the inertia provided by the storage of heat in solar thermal power.

- **Input variables**: The only exogenous variable that will be used in all models are the historical values of these capacity factors and temporal variables such as the month or hour. A purely time series approach has been taken, as opposed to other physical approaches that leverage weather forecasts. The intuition for this decision is that the information provided by the weather or any other significant variable should be integrated in the capacity factor, and relying on other external variables forces a two step approach where a model for those variables needs to be created and then a model for the capacity factor needs to be created, as there are no accurate forecasts for any of the other possible exogenous variables with a long enough timeframe.

- **Geographical scope**: The capacity factor will be that of the national Spanish power system.

- **Temporal scope**: The model should be able to forecast the capacity factor for a period of 1 to 10 years into the future, with hourly values.

- **Methodological scope**: Statistical, machine learning and deep learning models will be studied as candidates to modeling this variable.

# 2 Literature Review

## 2.1 Overview of renewable energy in Spain

## 2.2 Solar power generation

## 2.3 Wind power generation

## 2.4 Modeling techniques

## 2.5 Performance metrics for model evaluation

# 3 Data Description

## 3.1 Data Sources

The datasets needed for the exploration of the time series and the fitting of the models have been obtained from the ESIOS – System Operator's Electronic Information System. This is an online portal where the System Operator, the REE – Spanish Electrical Grid – uploads all relevant information for the operation of the grid.

In order to accomplish everything outlined in subsection 1.3 six different datasets are needed. These datasets are:

- Real time wind energy hourly generation
- Real time solar PV energy hourly generation
- Real time solar thermal energy hourly generation
- Wind energy monthly installed capacity
- Solar PV energy monthly installed capacity
- Solar thermal energy monthly installed capacity

## 3.2 Data Preprocessing

The obtained datasets are already very clean, so no missing or obviously wrong values were found on any of them. The installed capacity is obtained on a monthly basis as there is not any lower granularity in the ESIOS portal. With this datasets, the capacity factor is obtained as explained in equation (1) by dividing the generated electricity by the installed capacity times the period time frame. The time frame in this case is always one hour. Like this the hourly capacity factor for the Spanish electrical system for the three different technologies is obtained. Note that the data obtained spans from mid 2015 until end of 2023, since no previous data is available.

## 3.3 Exploratory data analysis

Now that it is well understood where the data comes from, an initial analysis of the data to understand what the capacity factor of each looks like will be performed.

### 3.3.1 Initial visualization

The first variable to be visualized will be that of the solar PV capacity factor, which can be seen in

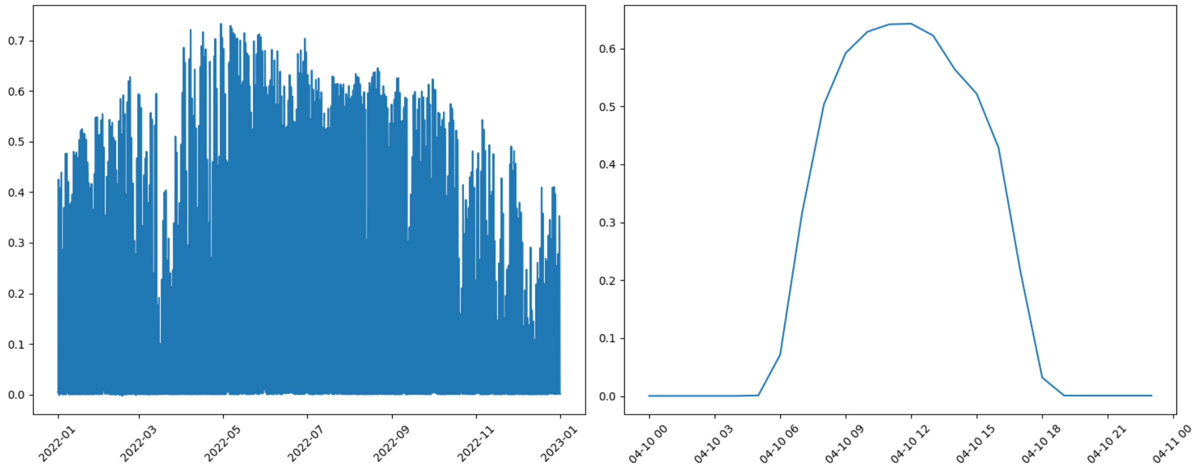


Figure 3: Yearly and daily profiles of the solar PV capacity factor, taken for 2022 and the $10^{\text{th}}$ April 2022 respectively.

The main features of the series can be well visualized here. The yearly seasonality is very clear, with higher peak values in summer and lower peak values in winter, although this is altered by periods of lower sunlight where peaks are reduced below the typical seasonal peak, like at the beginning of March. The daily pattern is also very clear, with values of practically zero – it is not exactly zero due to some background light and noise in the measurement – at night and a generation in accordance with the position of the sun in its ecliptic.

The solar thermal profile, although belonging to the same renewable resource as the solar PV, is slightly different.

Several differences between these profiles shown in Figure 4 and the previous ones are apparent. The yearly pattern, even though it follows roughly the same seasonality with higher peaks in summer is much more volatile than that of the solar PV. Furthermore, there are whole periods of several days where generation does not go to zero, due to the inertia of the thermal system. The highest peaks are also whigher for the solar thermal, reaching generation values

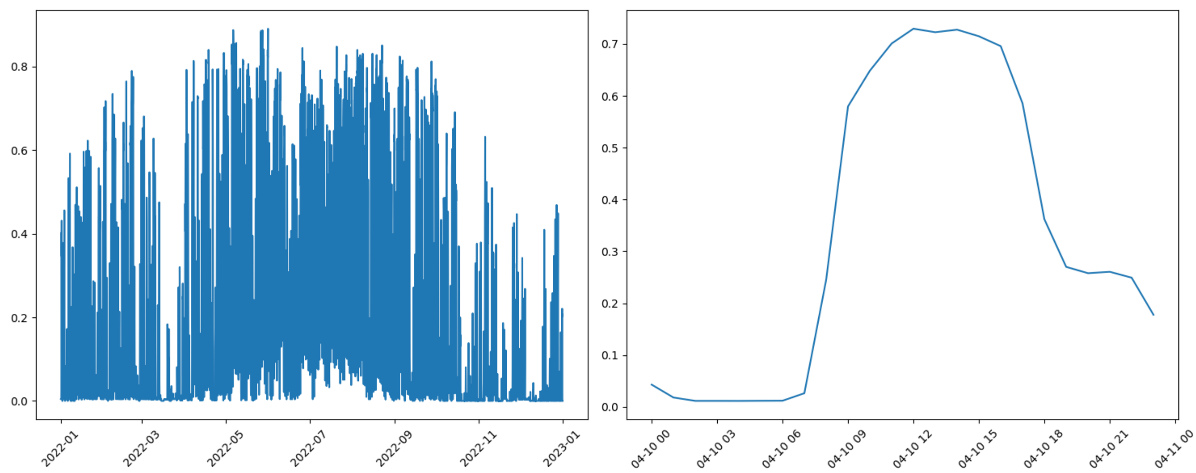


Figure 4: Yearly and daily profiles of the solar thermal capacity factor, taken for 2022 and the 10th April 2022 respectively.

much closer to its rated capacity in the summer than the solar PV. In the daily pattern the mentioned inertia can be clearly seen, where appart from the general sinusoidal throughout the day it can be seen how generation remains above zero at night.

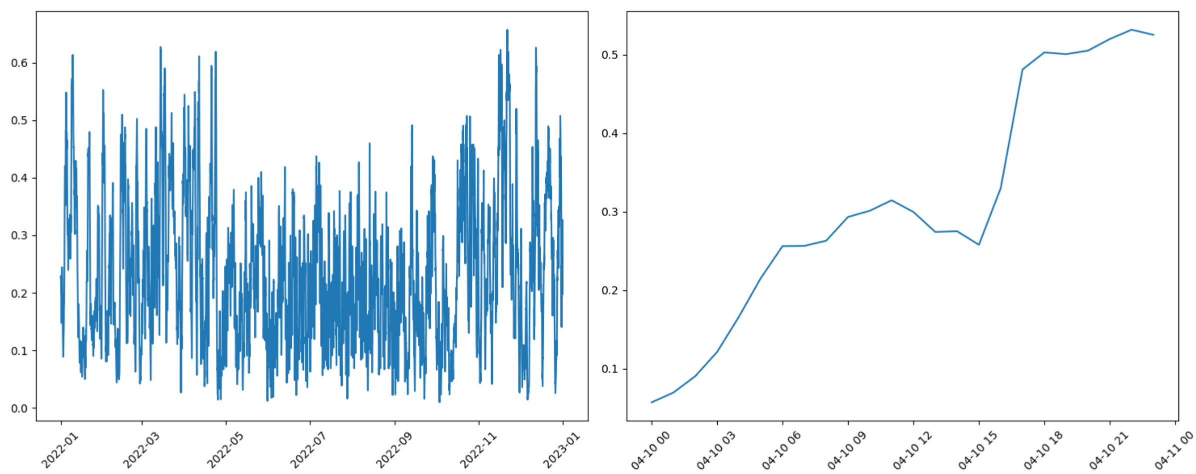


Figure 5: Yearly and daily profiles of the wind capacity factor, taken for 2022 and the 10th April 2022 respectively.

The wind profile shown in Figure 5 is the most different of the three. The yearly profile is the most volatile of the three, with less of a clear seasonal pattern although it can be seen how the mean values in summer seem to be lower than those in the winter. In the daily profile there are not any obvious patterns which can be seen at first view.

### 3.3.2 Individual statistical characteristics

After an initial view of the general time series for the three variables, a deeper look at the main statistical properties of these individual series is pertinent. For this deeper look a KDE and cumulative KDE will be shown and the main moments will be calculated.

The Kernel Density Estimate (KDE) is a non-parametric way of estimating the probability density function (PDF) of a random variable. It works by placing a kernel – that is a smooth, symmetric function – at each data point and then summing these kernels to produce a smooth estimate of the PDF. In more precise terms, the estimation of the PDF at $x$ denoted by $\hat{f}(x)$ is calculated as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2}$$

where $n$ is the number of datapoints, $K(\cdot)$ is the kernel function, $h$ is the bandwidth – a parameter used to approximate the width of the kernel and thus how smooth or sensitive to individual points the estimation is – and $x_i$ are the individual datapoints.

The chosen kernel function has been the Gaussian kernel, which is calculated as

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}x^2 \tag{3}$$

The method selected for estimating the optimal bandwidth has been the method outlined in Scott, 1979.

The cumulative KDE is an analogous method used for calculating the cumulative density function (CDF) instead of the PDF. It is calculated as

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{x} \frac{1}{h} K\left(\frac{t - x_i}{h}\right) dt \tag{4}$$

The KDE and cumulative KDE of all three variables can be seen in Figure 6

Several observations can be made regarding these plots. First of all, the high density of zero like values in both solar plots is clear, with an even higher denisty in the solar PV. In fact it can be seen how both KDE have non zero values for x values below zero, however this only happens due to the symmetry and width of the gaussian kernel, which for the points at zero sums some value for points below zero. Appart from the mode at zero, both plots have another peak near their maximum value, at around 0.6 and 0.9 for the solar PV and solar thermal respectively.

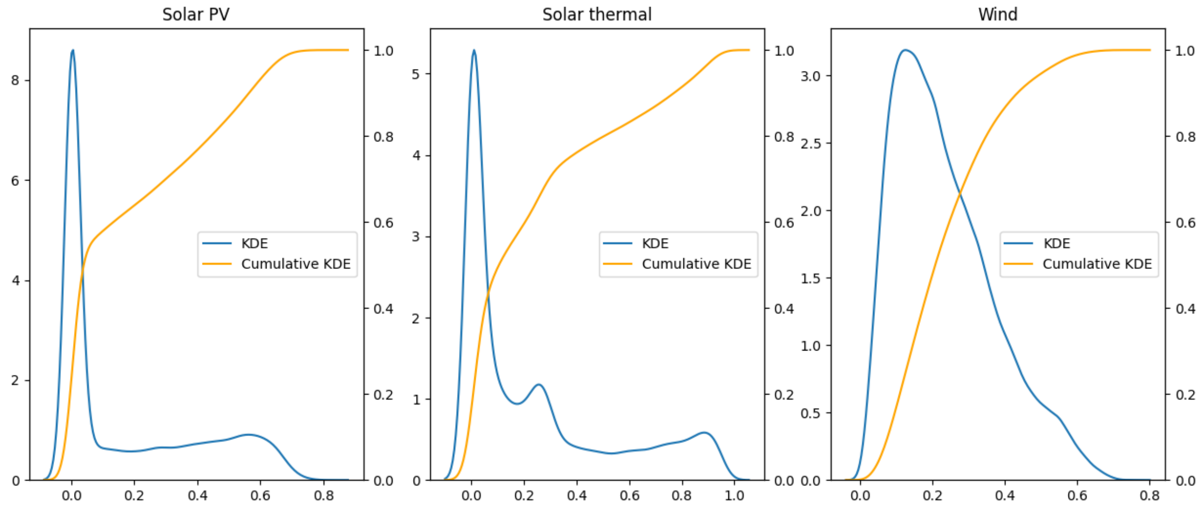

Figure 6: KDE and cumulative KDE of the three capacity factors

This is due to the shape of the solar ecliptic, where the sun remains at its highest point for longer than it is at any given point during ascent or descent. The solar thermal PDF also has a significant peak at around 0.3, probably due to this being the power at which the power plants runs with the stored heat at night. This is in fact the point where the daily curve in Figure 4 can be seen operating at night. As for the wind PDF and CDF, it is a much smoother plot, with the mode around 0.1 and then slowly and gradually tapers of.

In order to have a more quantitative view of these distributions, metrics relating to the first four moments are shown in Table 1

| Metric | Value | | |
|---|---|---|---|
| | Solar PV | Solar Thermal | Wind |
| Mean | 0.18 | 0.24 | 0.23 |
| Standard Deviation | 0.23 | 0.29 | 0.14 |
| Skew | 0.88 | 1.12 | 0.78 |
| Kurtosis | -0.79 | -0.08 | 0.01 |

Table 1: Main moments of the three capacity factors

Through these measures it can be seen how some of the intuitions obtained when looking at the time series plot of the different capacity factors were wrong. For example, the wind capacity factor is not the most volatile as it seemed, but in fact has the lowest standard deviation. This could be however due to the high variability of the solar profile, and taking away the seasonality of the three variables these metrics could be different.

### 3.3.3 Lagged relationships

Another key factor about the three series is their seasonality and how a given value at $t$ is influenced by past values at $t-1, t-2...$. In order to explore these relationships, two different metrics will be used.

The first one is the the autocorrelation. While correlation as expressed by the Pearson correlation coefficient is generally calculated like

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{5}$$

The autocorrelation function of a series quantifies the correlation of a series with lagged values of itself at different lags. For a given lag $k$ it is thus calculated like

$$ACF(k) = \frac{\sum_{t=1}^{N-k}(x_t-\bar{x})(x_{t+k}-\bar{x})}{\sum_{t=1}^{N}(x_t-\bar{x})^2} \tag{6}$$

The correlation coefficient has some known drawbacks however. It only measures linear relationships, so any dependence appart from this very simple relationship cannot be captured through a high correlation coefficient. That is what the new correlation coefficient proposed by Chatterjee, 2020 aims to solve, and why it has been included here. This new correlation coefficient, capable of capturing non linear relationships and with some very interesting statistical properties is calculated as

$$\xi(X,Y) = 1 - \frac{3\sum_{i=1}^{n-1}|r_{i+1}-r_i|}{n^2-1} \tag{7}$$

Where the given pairs $(x_1,y_1),(x_2,y_2)...(x_n,y_n)$ have been aranged so that $x_1 \leq ... \leq x_n$. Here $r_i$ denotes the rank of $y_i$. That is, the number of $j$ such that $y_j \leq y_i$. If there are ties among the $x$ values, the formula is slightly different, however with the capacity factor being a continuous variable it can be assumed that no ties will be present. Finally, note how even though $\rho \in [-1,1]$ the new correlation coefficient makes no distinction between positive and negative relationships and $\xi \in [0,1]$.

With this new correlation coefficient, a new autocorrelation function is constructed in the same way as in (6), calculating the correlation of the series with its lagged values for each lag. The following figure shows both autocorrelation functions for the three series.

Several observations can be made from the two autocorrelation functions in Figure 7. The

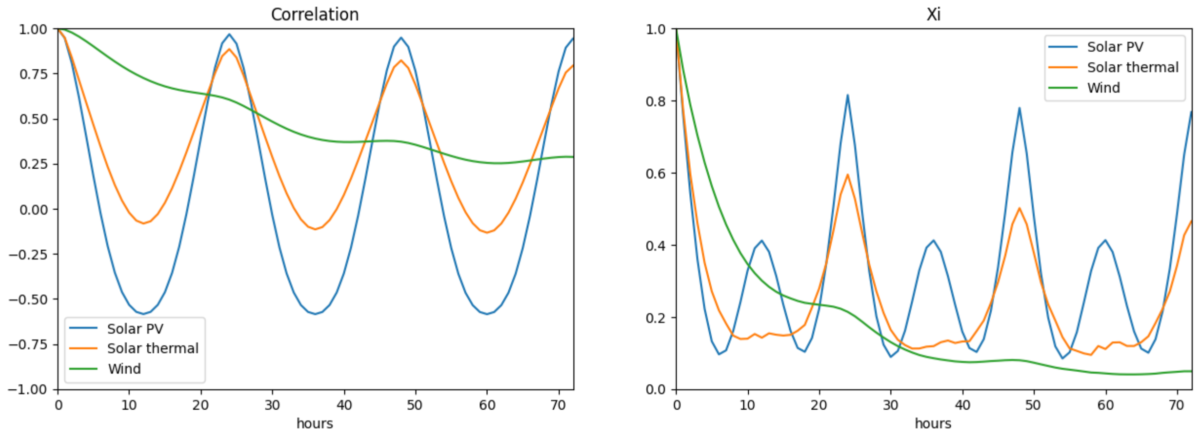


Figure 7: Autocorrelation functions with the two correlation coefficients for the three capacity factors at different hourly lags

seasonality in both solar series is again made apparent, with high autocorrelations at the 24, 48 and 72 hour mark showcasing the importance of the previous days' values at the same hour for a given day, but also at the 12, 36 and 60 hour mark. This second autocorrelation, negative in the case of pearson correlation, denotes the relationship between opposite moments of the day between different days. It is also noted how the correlation in the case of the solar PV is much more persistent, with the correlation peaks remaining more or less constant for the three days, while for the solar thermal this correlation tapers off a little bit faster. It is also interesting to see how the correlation at the 12, 36 and 60 hour mark is a lot lower for the solar thermal, showcasing how this relatinoship between values at a lag of 12 hours practically – although not completely – disappears, again due to the inertia of the thermal system which makes the solar thermal cycle less related to the sun's trayectory. As for the wind autocorrelation functions, it is much more linear and smooth, with the relevance of past values reducing gradually as the hours go by. There is a slight daily seasonality as well with the autocorrealtion at daily lags being slightly above what could be expected from the linear trend. However, this has a practically negligible relevance compared with the solar series. In fact, for the wind capacity factor it can be seen how the immediately previous values have much higher significance, while previous lags at the daily or semidaily periods hold much less significance.

INCLUDE SEASONAL LAGS?

### 3.3.4 Seasonal statistical characteristics

Even though the characteristics shown in previous sections already provides very significant information, it seems like many of these characteristics may be seasonally dependent. That is, they may not be the same in summer and in winter, or at night and during the day. That is why the previous analyses will be performed again but seasonally separating the datasets. Both daily and yearly splits will be done, where "day" is the period from 06:00 to 18:00 UTC and "night" is the period from 18:00 to 06:00 UTC. On the yearly split and taking into account the day's number within a year, spring goes from the 3$^{\text{rd}}$ to the 5$^{\text{th}}$ month, summer from the 6$^{\text{th}}$ to the 8$^{\text{th}}$, autoumn from the 9$^{\text{th}}$ to the 11$^{\text{th}}$ and winter from the 12$^{\text{th}}$ to the 2$^{\text{nd}}$, both inclusive in all cases. Note how these are metereological seasons, and not astronomical seasons based on the solstices and equinoxes.

The yearly seasonal breakdown can be seen in Figure 8.

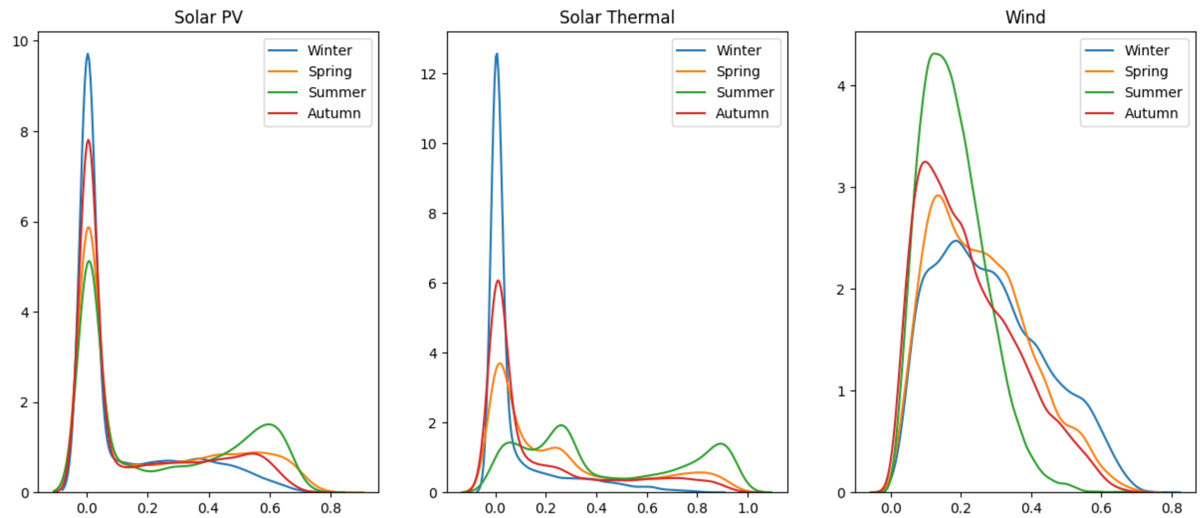


Figure 8: KDE plot for the three capacity factors for the different yearly seasons

For the solar PV it can be seen how on one extreme in winter the concentration of zeroes is highest, with very low density in the previously seen peak at 0.6, while in the summer the concentration of zeroes is still high – due to the night hours – but much less, and the peak at 0.6 can be clearly seen. Spring and autumn are found in between these values, with spring having a density of zeroes slightly above summer but a much lower peak at 0.6. For solar thermal the difference in density at zero is much higher, with winter having higher density and the rest of the seasons having a much lower one. In fact, for summer density at zero es very low, much more than other peaks. The peak at 0.9 can again only be seen in the summer, and the peak at 0.3 can be seen mainly for summer but also slightly for spring. As for the wind distribution, the summer season seems to be the one with the lowest volatility and lowest mean, while winter has the widest PDF and apparently mean.

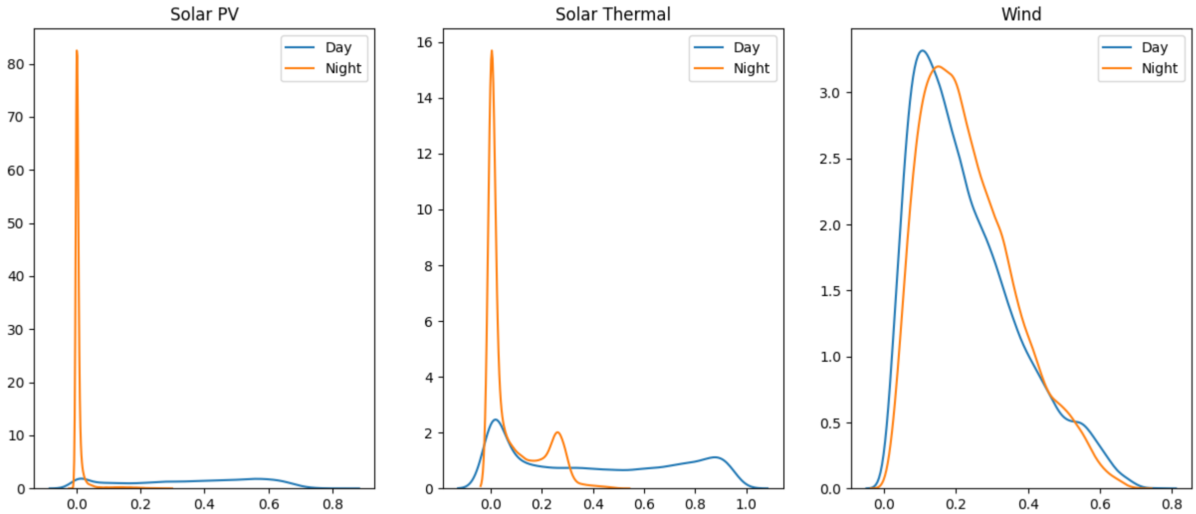


Figure 9: KDE plot for the three capacity factors for the different daily periods

In Figure 9 a similar representation for the KDE can be seen but for the split between night and day. For solar PV the plot is not very interesting, with the night having all its density at the zero point, with the day density having a slight peak at zero – as the exact times for sunset and sunrise vary during the year and the period defined here as "day" can in fact have some night time – and the already known one at 0.6. For solar thermal it can be seen how the peak at 0.9 is exclusively a day time fenomenon while the peak at 0.3 is a night time fenomenon, confirming the hypothesis that this was the power generated at night with the stored heat from the day. For the wind, it can be seen how the difference between day and night is practically negligible, with day exhibiting a slightly more "summer" like behaviour – that is, slightly less volatile and lower mean.

### 3.3.5 Multivariate analysis

Up until this point every analysis has been performed on a per variable basis, examining their individual properties. However, when thinking about the true implications for the power grid it becomes clear that the joint characteristics are also fundamental. For a given wind and solar distributions, the consequences of the grid are not the same if peaks in both happen at the same time or if they complement each other. That is why the joint characteristics will be studied, and they will mainly be studied through cross correlation and through copula analysis. Some other analysis, like through Granger causality as outlined in Granger, 1969 were attempted, but no new relevant information that wasn't uncovered through cross correlation was found.

Cross correlation follows a principle very similar to the autocorrelation, but instead of measuring the correlation of a series with lagged versions of itself, the correlation of a variable with lagged versions of a different variable are measured. That way, something similar to the predictive value of the second variable on the first can be understood. Again, the two different correlation coefficients already explained were used. Note how for Person's correlation the relationship for $(X, Y)$ is the same as for $(Y, X)$. However, when using lagged values the correlation does depend on which variable is lagged. for the new correlation coefficient the directionality matters, as the effect of one variable on a second is not the same as that of the second on the first one.

In Figure 10 the lagged cross correlation between each pair of variables can be seen. For each of the three variables, it is shown how the lagged versions of the other two are correlated to the given variable, for each of the two correlation coefficients.

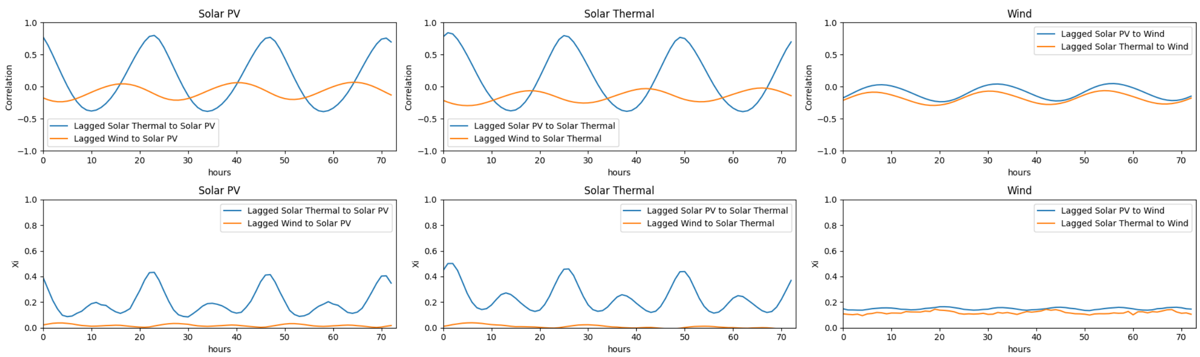


Figure 10: Lagged cross correlation between each pair of capacity factors

The similarity between the solar PV and solar thermal is clearly shown, with high correlations in both directions. Interestingly, the solar thermal capacity factor is more highly correlated with the 1-lag solar PV than with the simultaneous solar PV, probably due to the time it takes for the working fluid to be heated up, while the PV cells are able to convert that sunlight into power instantaneously. The second notable observation is that the predictive power of solar on wind seems to be higher than that of wind on any solar when looking at the Xi. Also, that correlation of the two lagged solar variables does not seem to decay for the examined 3 day period, and a very slight lag effect is seen where the daily seasonality of the sunlight is seen in the correlation, with wind being negatively correlated to simultaneous solar power but positively correlated to solar power at a lag of around 8 hours.

INCLUDE SEASONAL CROSS LAGS?

The second method used to see how the three variables compare to each other will be using copulas. Copulas are multivariate functions used to model the dependence between several random variables. Consider the random vector $(X_1, X_2, X_3)$ representing the observations of capacity factor for solar PV, solar thermal and wind energy for a given point in time. The marginal CDF of each of the variables is represented as $F_i(x) = P(X_i \leq x)$, which when applied to the random vector yields the marginals $(U_1, U_2, U_3) = (F_1(X_1), F_2(X_2), F_3(X_3)) \in [0,1]^3$. The copula $C(\cdot)$ is the joint cumulative distribution function such that

$$C(u_1, u_2, u_3) = P(U_1 \leq u_1, U_2 \leq u_2, U_3 \leq u_3) \tag{8}$$

The copula contains all information on the dependence structure between the three series, while the marginal cumulative distribution functions contain all inforamtion on the marginal distribution of the variables.

In this analysis, the marginal CDF used has been the gaussian cumulative KDE outlined in (4). The copula used has been a gaussian copula, which relies on the multivariate normal distribution.

For the fitting process, a cumulative gaussian KDE is fitted for each of the variables and then the univariate values are transformed according to that CDF. Then, the normal percent point function (PPF) – the inverse of the CDF – is applied to transformed variable to normalize it. Then, a multivariate gaussian is fitted on the normalized data.

The correlation matrix of the gaussian copula can be seen in Table 2.

| | Solar PV | Solar thermal | Wind |
|---|---|---|---|
| Solar PV | 1.00 | 0.72 | -0.19 |
| Solar thermal | 0.72 | 1.00 | -0.23 |
| Wind | -0.19 | -0.23 | 1.00 |

Table 2: Correlation matrix of the gaussian copula with gaussian KDE marginals

It can be seen how in the copula representation the strong positive correlation between both solar series and the negative correlation between wind and both solar series still stands.

# 4 Methodology

In this section several points regarding the methodology used in the work will be presented. The general approach taken to modeling, as well as the models themselves will be explained. The metrics used to evaluate the goodness of fit of the different models will also be presented.

## 4.1 Overview of modeling approach

In order to understand why the models that will be presented below have been chosen and why the evaluation metrics have been chosen, it is worth first taking a look at the requirements of the modeling task and the general criteria that determine what a good result is.

### 4.1.1 Modeling task characteristics

The general technical characteristics of the task, from the objectives and the time series characteristics already explored are the following:

- **Modeling horizon**: The horizon to be modeled is of between 1 and 10 years, or between 8.760 and 87.600 hourly timesteps.
- **Available data**: The data available for the training and testing tasks consists of slightly below 10 years of data, or slightly below 87.600 observations. This data has been consistently recorded at a 1 hour timestep interval.
- **Seasonality**: There has been found to be high seasonality, both at the daily and yearly frequencies.
- **Exogenous variables**: A certain predictive power of some series with respect to the others has been found. Thus the capacity of incorporating the use of exogenous variables or multivariate modeling aproaches are specially interesting.
- **Extreme values**: There is a special interest in having accurate modeling of the extreme values, specially in the wind capacity factor.
- **Pointwise forecast**: The forecasts produced by the model can be pointwise predictions, there is no need to produce probabilistic forecasts for each timestep.

### 4.1.2 Modeling criteria

Starting with the objectives already outlined in the section called Objective and scope, here these objectives will be broken down into more specific criteria that the best model must aim to meet in the best possible way. The specific characteristics the model must meet are the following:

- **Seasonality**: The chosen models must be able of incorporating seasonality into its forecast.

- **Multiple variables**: The chosen models should be able to incorporate multiple variables, be it through a multivariate modeling, through historical exogenous variables and/or through future exogenous variables. Through the modeling task it will be made apparent wether incorporating the information of the other series helps in modeling a given series.

- **Complex relationships**: The temporal relationships of a series with itself and the relationships of a series with the other variables are probably more complex than linear relationships. That is why a model capable of incorporating complex non linear temporal and intervariate relationships will probably have an advantage.

- **Multiperiod modeling**: As it has already been mentioned, the modeling period will encompass between 8.760 and 87.600 timesteps. The model must be able to create such forecasts. But going further, if the prediction is created by autoregressively estimating the next step, there is a compounding in errors over time Selim et al., 2020. By estimating multiple timesteps in one prediction pass, the number of autoregressive steps can be reduced and with it the compounding of errors can be reduced too. Therefore, models with bigger single-step prediction window should have an advantage.

- **Long and short term temporal relationships**: The model must be able to incorporate both long term and short term temporal relationships.

- **Robustness**: The model must be resistant to overfitting, given the parameter size of some of the models and the number of available training datapoints.

- **Extreme value accuracy**: The model must be specially sensitive to extreme values, be it through the flexibility of incorporating a custom loss function or through its intrinsic characteristics.

- **Computational efficiency**: Given the long modeling horizon, the model must be efficient both in prediction time and in memory used when training and when incorporating historical data in prediction.

## 4.2 Models

Now that the criteria used to select potential models has been outlined, the different models that will be evaluated can be presented. These models have been divided by model family or type. Here below, the different model types and each of the models and their characteristics and main working principles are explained.

### 4.2.1 Statistical models

These models are based on classical statistical methods to analyze and predict temporal series. They use techniques like regression analysis, autocorrelation and spectral analysis to model and forecast temporal variables. The ones that will be used here are:

- **Linear Regression**: This is the most simple model and it will be used as a naive benchmark. It models a given capacity factor value for one of the series as a linear combination of several exogenous variables. In this case, those variables are the values of fourier sine and cosine terms, some lagged values of the given time series and some lagged values of other time series, which in this case will be the other capacity factors. The main parameters of the model are thus:
  - Seasonal frequencies: That is the frequencies for which the fourier terms will be calculated. It can be daily, yearly, both, etc.
  - Fourier harmonics: For each of the seasonalities, several harmonics of fourier terms can be calculated, creating more nuanced patterns the more harmonics are added.
  - Autoregressive order: That is, what lags of the modeled series to incorporate.
  - Exogenous lags: That is, what lags of the other two series to incorporate.

The mathematical formulation for the model thus is:

$$Y_t = \beta_0 + \sum_{\tau}\sum_{k}\left[\gamma_{\tau,k}\sin\left(\frac{2\pi kt}{\tau}\right) + \delta_{\tau,k}\cos\left(\frac{2\pi kt}{\tau}\right)\right] + \\ + \sum_{l}\phi_l Y_{t-l} + \sum_{v}\sum_{p}\theta_{v,p}X_{v,t-p} + \epsilon_t \quad (9)$$

where $\tau$ is each of the seasonality periods, $k$ is each of the fourier harmonics, $l$ is each of the self lags, $v$ is each of the other two capacity factors and $p$ is each of the lags used for the two other capacity factors.

This model opperates on several assumptions which are worth mentioning. The first is that all relationships – those with the seasonal terms, with lagged values, with the other capacity factors, etc. – are assumed to be linear. The error term $\epsilon_t$ is assumed to be independent and identically distributed (i.i.d.) with constant variance (homocedastic) and gaussian, although there is no evidence that this is true.

- **ARIMAX**: The AutoRegressive Integrated Moving Average with eXogenous variables model is an extension to the previous model. It incorporates both autoregressive and moving average components to the endogenous time series and allows for time integration, which extends the purely autoregressive view in the previous model, as well as some exogneous variables. It provides a more complex representation of the relationship of the time series with its lagged values than the linear regression model presented above. To the previously presented parameter, these new ones are added:

  - Integration order: This represents the number of times the endogenous time series is differentiated.
  - Moving average order: That is, what moving average error terms are incorporated.

  The new model formulation is:

$$Y_t = -\left(\Delta^d Y_t - Y_t\right) + \beta_0 + \sum_{\tau}\sum_{k}\left[\gamma_{\tau,k}\sin\left(\frac{2\pi kt}{\tau}\right) + \delta_{\tau,k}\cos\left(\frac{2\pi kt}{\tau}\right)\right] + \\ + \sum_{l}\phi_l \Delta^d Y_{t-l} + \sum_{q}\theta_q\epsilon_{t-q} + \sum_{v}\sum_{p}\theta_{v,p}X_{v,t-p} + \epsilon_t \quad (10)$$

  where $d$ is the order of differentiation with $\Delta Y_t = Y_t - Y_{t-1}$ and $q$ is lags of the error term added for the moving average.

- **VARMAX**: The Vector AutoRegressive Moving Average with eXogenous variables is a modification of the ARIMAX model, but without the integration component and turning the single variable endogenous variable into a vector variable. This model jointly predicts all capacity factor variables, directly creating the joint distribution. The model is formulated as:

$$Y_t = \beta_0 + \sum_{\tau}\sum_{k}\left[\gamma_{\tau,k}\sin\left(\frac{2\pi kt}{\tau}\right) + \delta_{\tau,k}\cos\left(\frac{2\pi kt}{\tau}\right)\right] + \\ + \sum_{l}\phi_l Y_{t-l} + \sum_{q}\theta_q \epsilon_{t-q} + \epsilon_t \quad (11)$$

  where $Y_t$ here is not a single variable datapoint but a three dimensional vector containing all three capacity factors. The coefficients are also vector coefficients.

### 4.2.2 Machine Learning models

Machine learning models are a class of models designed to learn more complex patterns from data with less constraining assumptions and generally more parameters. Even if these models are quite general and are not specifically designed for time series modeling, they can be readily used for this purpose. For this specific use case, these models excel at capturing complex and non linear relationships that statistical methods may overlook, they can therefore outperformed the former methods when these characteristics are present or when data is highly non-stationary or with intricate dependencies. The ones that have been implemented are:

- **SVM**: Support Vector Machines introduced in Cortes and Vapnik, 1995 are ML models that can be used for either classification or regression. They work by finding the optimal hyperplane that separates the datapoints into distinct classes or predicts values with the maximum margin. In the specific case of regression the model is known as Support Vector Regression (SVR). Its objective is to find a function $f(x)$ that deviates from the empirical values $y_i$ by at most a margin $\epsilon$ ensuring the model remains as flat as possible. That is, preventing overfitting. The function takes the form:

$$f(x) = w^T\phi(x) + b \quad (12)$$

where $w$ is the weight vector, $b$ is the bias term and $\phi(x)$ is a transformation that maps the input data into a higher-dimensional space. The SVR algorithm solves the following optimization problem:

$$
\begin{aligned}
\min_{w,\xi_i,\xi_i^*} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\
\text{s.t.} \quad & y_i - f(x_i) \leq \epsilon + \xi_i \\
& f(x_i) - y_i \leq \epsilon + \xi_i^* \\
& \xi_i, \xi_i^* \geq 0
\end{aligned} \tag{13}
$$

where $\xi_i$ and $\xi_i^*$ are slack terms allowing for some flexibility in prediction errors and $C$ is the regularization parameter that controls the tradeoff between the flatness of the function and how well it fits the data. Solving the optimization problem provides:

$$
f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\, K(x_i, x) + b \tag{14}
$$

where $\alpha_i$ and $\alpha_i^*$ are the optimal lagrange multipliers corresponding to the slack variables $\xi_i$ and $\xi_i^*$ and $K(\cdot)$ is the Kernel function used to calculate the distance between predicted and empirical datapoints and which provides the needed non linearities to the function. In this case, the Radial Basis Function has been used with

$$
K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2} \tag{15}
$$

where $\gamma$ controls the spread of the kernel.

As it can be seen, the SVR is quite useful as it includes non-linear relationships, has built in robustness to overfitting and is quite scalable. However, it is quite sensitive to the choice of parameters – mainly that of kernel, $C$ and $\epsilon$ – and assumes the relationship between input space and prediction remains constant in time, so it is not that appropriate for non stationary time series.

In the implementation used in this work, the input space is equivalent to that of the classical statistical models, comprised of some lags of the data, some fourier decompositions at different levels and some lags of the other time series.

- **XGBoost**: eXtreme Gradient Boosting is a highly efficient and scalable implementation within the paradigm of decision trees. Decision trees are another ML framework where starting from a root node, the dataset or output space is split into two or more subsets based on a specific value of a selected feature. For example, a very simple example would be to say that on average the predicted solar capacity factor is 0.3 (root node). If it is daytime the predicted value is 0.6 and if it is nightime the predicted value is 0.0 (two subsequent nodes). This split can be further refined with more levels on different values of different features, providing a more accurate and non linear prediction based on the input values. Trees are built through an algorithm which at each node selects the feature that best separates the data based on a criterion such as entropy. This model is quite simple but allows for non-linearity and is very robust to different scales in input features. However, it can easily overfit on the training data and the results can be quite instable.

Returning to XGBoost, this model builds sequential trees where each new tree corrects the errors of the previous one in what is known as an ensemble of trees. It also implements several improvements such as regularization, tree pruning or parallelization. The model thus builds a prediction like

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{16}$$

where $\hat{y}_i^{(t)}$ is the prediction after $t$ trees and $f_t(\cdot)$ is the tree added at level $t$. Each individual tree is constructed optimizing an objective function which both minimizes the loss function between true and predicted value and also has a regularization term which prevents overfitting. The objective thus becomes

$$\min_\theta \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{17}$$

with the regularization term for each tree $k$ being

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{18}$$

with $T$ being the number of nodes in the tree, $w_j$ being the weight of leaf $j$ and $\gamma$ and $\lambda$ being regularization parameters.

This model has some of the same advantages like the ability of capturing non linear relationships and built in regularization, but it also allows for some features like custom loss functions.

In the implementation used for the XGboost, the input space is again equivalent to that of the classical statistical models with some lags of the data, some fourier decompositions at different levels and some lags of the other time series.

### 4.2.3 Neural Network models

### 4.2.4 Transformer models

## 4.3 Model training and validation

## 4.4 Evaluation metrics

In order to understand why the following metrics have been chosen for evaluation it is important to keep in mind the overarching goal of the project. The goal of the mentioned models is not to provide an accurate short term forecast, but to provide a model that porperly simulates the long term dynamics of the three time series. That is, models that are able of producing long term scenarios of solar and wind generation with univariate and multivariate dynamics equivalent to the real world ones, in order to optimize and test several grid planning possiblities on these scenarios.

In order to best achieve this goal, the best performing model will have to be evaluated on different metrics, each corresponding to different attributes needed for correct whole picture modeling. These are the different attributes that will be tested and the metrics that will be used to test them.

### 4.4.1 Marginal distributions

The first metric that will be evaluated will be the similarity of the univariate series with the real world data. Note that this similarity is not a pointwise similarity. In many relevant papers, the accuracy of a series predicition is calculated through metrics like the RMSE or the MAPE. However, these metrics require that the prediction try to get right the value at each timestep. If there is a delay of one timestep, or a "cloudy week" comes a week earlier or later the RMSE will give a very poor score even though the general behaviour is the same. That is why the approach taken here will be slightly different. The metrics used will be the following:

- **Cramér von Mises criterion**: The Cramér von Mises criterion intriduced by Cramér, 1928 measures the distance between two CDF. It is calculated like:

$$\omega^2 = \int_{-\infty}^{\infty} \left[F_m(x) - F_e(x)\right]^2 dF_e(x) \tag{19}$$

Where $F_m(x)$ is the modelled empirical CDF and $F_e(x)$ is the observed empirical CDF.

- **Kullback-Leibler divergence**: The KL divergence or relative entropy introduced in Kullback and Leibler, 1951 is a statistical distance that measures how different a probability distribution is from another reference one. It is calculated like

$$D_{KL}(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{20}$$

where $P$ is the modelled probability distribution and Q is the reference probability distribution.

- **ACF distance**: Even though the marginal distributions are by themselves a fundamental part of the closeness of the univariate series to reality, they are not the only one. Another key aspect of the series is its termporal self dependency, studied through the ACF of the correlation coefficient and the new correlation coefficient in the section called Lagged relationships. The difference between the ACF of the $\xi$ metric of the empirical data and the model will be measured through a weighted root mean squared error, calculated as:

$$\text{ACFD} = \sqrt{\frac{\sum_{k=0}^{N} w_k \left[ACF_m(k) - ACF_e(k)\right]^2}{\sum_{k=0}^{N} w_k}} \tag{21}$$

where $N$ is the number of periods for which the metric is calculated – 72 in this case, $ACF_m$ is the model's $\xi$ ACF, $ACF_e$ is the empirical $\xi$ ACF and $w$ is the weighting factor, determined to be $w_k = ACF_e(k)$. This has been chosen so a higher weight is given to those lags with a higher empirical relevance.

- **Seasonal metrics**: The metrics above will be calculated on the overall sample but also on subsamples divided by yearly season and daily period, to ensure the fit not only on a global scale but also on specific periods.

### 4.4.2 Joint distributions

Now that it is clear how the marginal distributions will be assessed, it is time to look at how the dependence structure between the three series will be tested. There are several metrics that will be used for that purpose:

- **Copula correlation matrix distance**: In the section regarding Multivariate analysis, the correlation matrix of a gaussian copula representing the dependence structure between the series was calculated. The same matrix will be calculated for samples of each of the models. Then, the distance (CCMD) between both matrices will be calculated. This distance is calculated as:

$$CCMD = \sqrt{\sum_i \sum_j (a_{ij} - b_{ij})^2} \tag{22}$$

  where $A = \{a_{ij}\}$ is the correlation matrix of the copula fitted to the modelled data and $B = \{b_{ij}\}$ is the correlation matrix shown in Table 2.

- **CCF distance**: The cross correlation function distance is analogous to the autocorrelation function distance explained above, but for the $\xi$ cross correlation function instead of ACF.

- **Seasonal metrics**: Similarly to the univariate case, the metrics above will be calculated on the whole period but also with seasonal differences and daily period differences.

### 4.4.3 Extreme value analysis

A significantly important aspect of the models is their need to be particularly accurate on the extreme values. Periods of extreme wind or solar generation are particularly interesting to the grid, due to their displacement of other energy sources or their need of them. That is why the accuracy of the models on extreme values will be assessed through these metrics:

- **Conditional Value at Risk**: The CVaR, usually used in financial risk management, measures what a certain value on average will be given that a certain threshold has been exceeded. It is calculated as

$$\mathrm{CVaR}_\alpha(x) = \mathbb{E}\left[X \mid X > \mathrm{VaR}_\alpha\right] \tag{23}$$

with $\mathrm{VaR}_\alpha$ being the Value at Risk for a given certainty. That is, the maximum value not exceeded with a probability $\alpha$. It is characterized as

$$\mathrm{VaR}_\alpha = \inf\{x|F_X(x) \geq \alpha\} \tag{24}$$

The assessment metric will be the difference between the CVaR calculated as

$$\mathrm{CVRD} = \frac{\mathrm{CVaR}^m_{95\%}}{\mathrm{CVaR}^e_{95\%}} - 1 \tag{25}$$

where $\mathrm{CVaR}^m_{95\%}$ is the CVaR at a 95% level for the model and $\mathrm{CVaR}^e_{95\%}$ is the empirical CVaR at a 95% level.

- **Tail dependence**: The tail dependence coefficient is a measure of the comovements of the tails of their distributions. It can be lower tail dependence or upper tail dependence, with the lower tail dependence calculated as

$$\lambda_l = \lim_{q\to 0} P\left(X_m \leq F_m^{-1}(q) \mid X_e \leq F_e^{-1}(q)\right) \tag{26}$$

and the upper tail dependence coefficient being calculated as

$$\lambda_u = \lim_{q\to 1} P\left(X_m > F_m^{-1}(q) \mid X_e > F_e^{-1}(q)\right) \tag{27}$$

with $F_m^{-1}$ being the inverse CDF for the model and $F_e^{-1}$ being the empirical inverse CDF. There will be two comparisons. The comparison of the tail dependence between each pair of modeled variables with that of the corresponding pair of empirical variables – e.g. tail dependence between modeled solar PV and modeled wind and tail dependence between empirical solar PV and empirical wind – and also the tail dependence between each modeled variable with its corresponding empirical varaiable – e.g. tail dependence between modeled solar PV with empirical solar PV.

- **Return level**: The GEV function is a function often used to model the maxima of sequences of random variables. Its CDF is

$$P\left(GEV\left(\mu, \sigma, \xi\right) \leq x\right) = e^{-t(x)} \tag{28}$$

with

$$t(x) \equiv \begin{cases} \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ e^{\left(-\frac{x-\mu}{\sigma}\right)}, & \text{if } \xi = 0. \end{cases} \tag{29}$$

As it can be seen from the formulation above the distribution of maximum values depends on three variables, with $\xi$ governing the tail behaviour. Once the GEV distribution is fitted to the data, the return level can be estimated. The return level is the value expected to be exceeded on average once in a given period. If the 100 year return level of wind capacity factor is 0.95, that means that on average there will be one instance every 100 years where the capacity factor exceeds 0.95. The return level $z_T$ is calcualted as

$$z_T = \mu + \frac{\sigma}{\xi}\left[\left(-\log\left(1-\frac{1}{T}\right)\right)^{-\xi} - 1\right] \tag{30}$$

Thus, the weekly maxima will be obtained with the data, with which a GEV distribution will be fitted and a 10 year return level will be calculated. The assessment metric will be the percentage difference in return level calculated as

$$\text{RLD} = \frac{z_{10}^m}{z_{10}^e} - 1 \tag{31}$$

This will be calculated for both the maxima and the minima of the distributions.

The extreme value assessment is particularly important for wind generation, rather than for solar generation as it is much more consistent. Therefore, these values will only be calculated for the wind series.

# 5 Results

## 5.1 Model performance

## 5.2 Discussion of results

# 6 Conclusion

Outline with conribution, summary of findings, limitations, future work

# Appendices

## A Code implementation

The complete code used in this work has been made publicly accessible. That is, the implementation of the different models, but also all the code use to train and evaluate them, the preliminary study of the datasets, etc. Even the latex version of this written document has been made accessible. All that code can be accessed via GitHub at https://github.com/fcelya/solar-wind-generation. The only elements that have not been made accessible are the datasets used to run the strategy and to estimate the factor parameters. They could not be uploaded due to GitHub's limitations, and enough information has been given in section 3 for anyone to be able to recreate those datasets.

The hope is that by making this code easily accessible, more people will be tempted to explore it, play with it, tackle the problem of long term solar and wind energy modeling and the ideas presented in this work can be further explored by the community at large.

# References


Agency, I. E. (2021). *Spain 2021 energy policy review*. https://iea.blob.core.windows.net/assets/2f405ae0-4617-4e16-884c-7956d1945f64/Spain2021.pdf

Ahmed, S. D., Al-Ismail, F. S. M., Shafiullah, Al-Sulaiman, F. A., & El-Amin, I. M. (2020). Grid integration challenges of wind energy: A review. *IEEE Access*, *8*, 10857–10878. https://doi.org/https://doi.org/10.1109/access.2020.2964896

Chatterjee, S. (2020). A new coefficient of correlation. *Journal of the American Statistical Association*. https://doi.org/https://doi.org/10.1080//01621459.2020.1758115

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/https://doi.org/10.1007/bf00994018

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, *1928*(1), 13–74. https://doi.org/10.1080/03461238.1928.10416862

Einstein, A. (1905). Über einen die erzeugung und verwandlung des lichtes betreffenden heuristischen gesichtspunkt. *Annalen der Physik*, *322*(6), 132–148. https://doi.org/https://doi.org/10.1002/andp.19053220607

Ellabban, O., Abu-Rub, H., & Blaabjerg, F. (2014). Renewable energy resources: Current status, future prospects and their enabling technology. *Renewable and Sustainable Energy Reviews*, *39*, 748–764. https://doi.org/https://doi.org/10.1016/j.rser.2014.07.113

Fernández, L. (2024). https://www.statista.com/statistics/1253518/clean-energy-investments-spain/

Glenk, G., & Reichelstein, S. (2022). The economic dynamics of competing power generation sources. *Renewable and Sustainable Energy Reviews*, *168*, 112758–112758. https://doi.org/https://doi.org/10.1016/j.rser.2022.112758

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438. https://www.jstor.org/stable/1912791

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/https://doi.org/10.1214/aoms/1177729694

Kumar, V., Pandey, A. S., & Sinha, S. K. (2016). Grid integration and power quality issues of wind and solar energy system: A review. https://doi.org/https://doi.org/10.1109/iceteeses.2016.7581355

Ministerio de Industria, E. y. T. (2015). https://www.boe.es/buscar/act.php?id=BOE-A-2015-10927

Ministerio de Industria, T. y. C. (2007). https://www.boe.es/buscar/act.php?id=BOE-A-2007-10556

Muljadi, E., Butterfield, C., Chacon, J., & Romanowitz, H. (2006). Power quality aspects in a wind power plant: Preprint. https://digital.library.unt.edu/ark:/67531/metadc878686/m2/1/high_res_d/875771.pdf

*Renewable electricity generation*. (2024). Red Eléctrica. https://www.sistemaelectrico-ree.es/en/spanish-electricity-system/generation/renewable-electricity-generation

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610. https://doi.org/https://doi.org/10.1093/biomet/66.3.605

Selim, M., Zhou, R., Feng, W., & Alam, O. (2020). Reducing error propagation for long term energy forecasting using multivariate prediction. In G. Lee & Y. Jin (Eds.), *Proceedings of 35th international conference on computers and their applications* (pp. 161–169). EasyChair. https://doi.org/10.29007/mbb7

Steen, D., Goop, J., Göransson, L., Nursbo, S., Brolin, M., Sarasini, S., & Kåberger, T. (2014). *9 challenges of integrating solar and wind into the electricity grid*. https://core.ac.uk/download/pdf/70609781.pdf

Vittal, V., & Ayyanar, R. (2013). Grid integration and dynamic impact of wind energy. *SpringerLink*. https://doi.org/https://doi.org/10.1007-978-1-4419-9323-6