Winter
2023/24

Time Series Analysis Course

Final Project

Tomer Hait

316350651

Tal Peer

208851253

Technion - Israel Institute of Technology

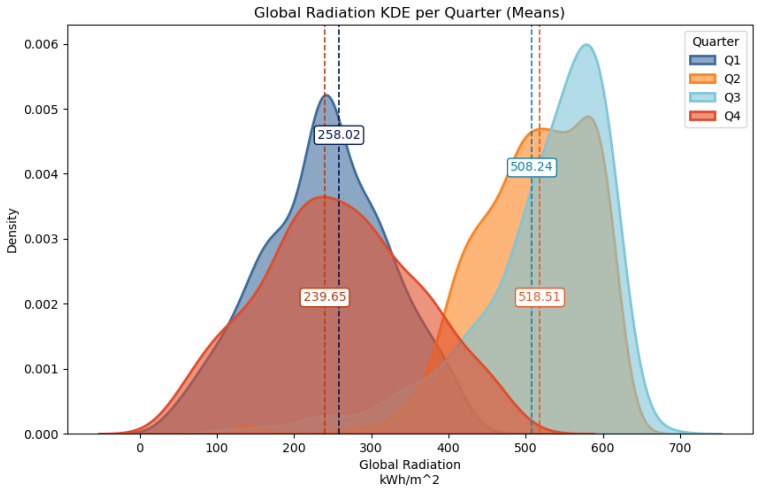
Table of Contents

Introduction

Our project will focus on the analysis of Global Radiation measurements between 2017-2023.

Global radiation refers to the total solar radiation received on a horizontal surface, encompassing both direct and diffuse solar radiation. This radiation is a key component of Earth's energy balance and plays a crucial role in various natural processes.

We selected a dataset spanning hourly measurements of global radiation recorded at the Haifa Technion station, covering the years from 2017 to 2023.

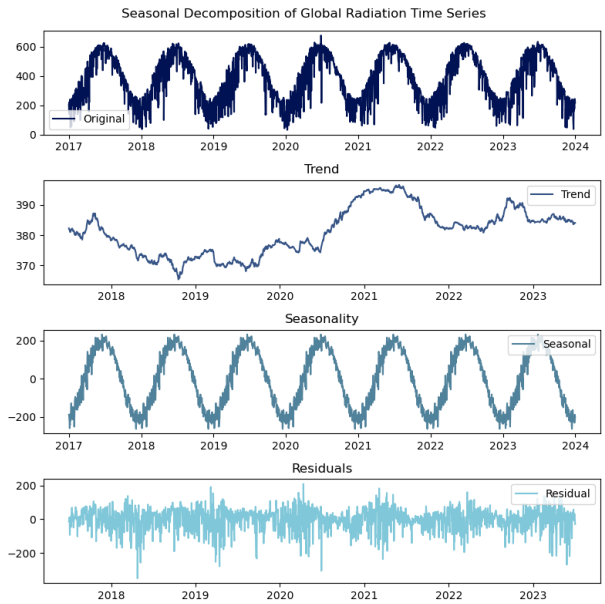


Originally, the dataset provided measurements for each hour from 8:00 to 20:00. However, we opted to aggregate the data to compute daily averages, along with additional temporal information.

The dataset was sourced from the Israeli Meteorological Service website, under the governance of the Ministry of Transport and Road Safety. It comprises 2546 records and the following attributes:

date	date of the recorded data and serves as the table index (each row corresponds to a specific date)
avg_daily_global_radiation	average daily global radiation measured on each date.
year	year of the recorded data.
month	month of the recorded data
day	day of the recorded data
quarter	categorizes the data into four quarters: Q1, Q2, Q3, and Q4. (manual mapping)

Decomposing the dataset, we can observe that seasonality manifests as fluctuations in global radiation levels throughout the year, with higher levels during certain seasons (e.g., Q3) and lower levels during others (e.g., Q1 and Q4).
We can observe an overall trend by increasing radiation levels over the mentioned years. Also, we can observe from the boxplot[1] and residuals plot a discernible regularities or structures by short-term fluctuations outliers in Q2.



More EDA is available in the Appendix section.

[1] Appendix 1

Methodologies

Model Fitting

While initially observing the data, we notice significant yearly seasonality. This observation is not surprising as radiation depends on the sun, which revolves around the earth based on a roughly estimated 365-days period. Thus, the first model which we considered is the SARIMA model.

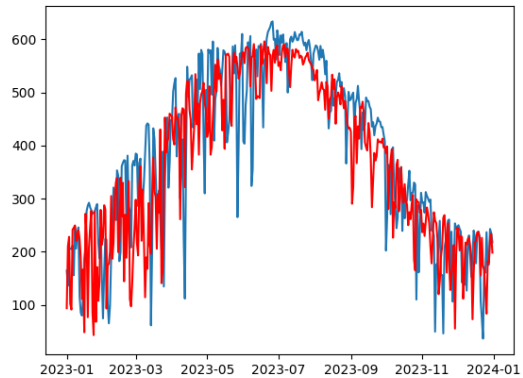
Note: In all our models, since the data is that of several years, we decided to predict the last year in the dataset based on the last one that was documented

SARIMA

Basic Fitting

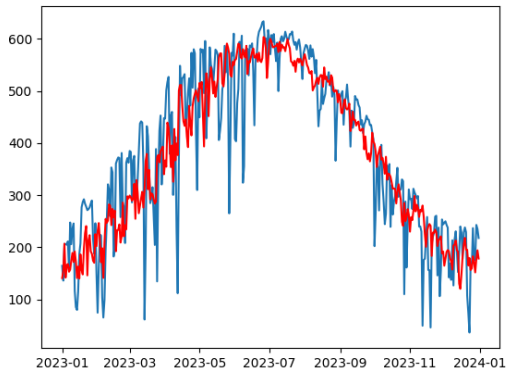
Viewing the ACF and PACF plots, we saw strong signals of autocorrelation.
The ACF plot resembled a decaying cosine wave. In addition, we did not see any seasonality in the PACF plot.
Therefore, we thought an appropriate model would be $\boldsymbol{SARIMA(0,0,1)(1,0,0,365)}$.
We set the differentiation parameters to 0 on both the seasonal and non-seasonal parts of the model, as we did not see any significant consistent trend in the decomposition.

Note: the time it took for the model to fit was very long. This is likely since the seasonality was high.



2023 Prediction by - SARIMA(0,0,1)(1,0,0,365)
RMSE 91.21

SARIMA works well when seasonalities are rather small.
A prediction over average monthly radiation was possible to reduce the run times, but we didn't find much motivation behind that process, and in addition, it would result in less observation than which were required in this task (less than 150). This made it difficult for us to try more SARIMA variations, due to run time and computer memory limitations. The **RMSE** of the SARIMA model was **91.21.**
In attempt to improve the RMSE on the test set, we have decided to apply a small grid search, and the model which behaved best was $\boldsymbol{SARIMA(0,0,0)(1,0,1,365)}$ with **RMSE** of **75.43** on the test set.



2023 Prediction by - SARIMA(0,0,0)(1,0,1,365)
RMSE 75.43

Note: Although it appears that taking a non-seasonal AR component might contribute to the model, it did the opposite - making the predictions rely solely on it instead of combining it with the seasonal AR component.

ARIMAX - Usage of Interior Properties as Exogenous Variables

In general, the use of basic SARIMA models was impractical due to extremely long run times.
However, as we saw in the lectures, there are other ways to model seasonality in time series analysis such as Fourier transformation.
A suggestion that was presented in the lecture was to include the modeling of the seasonality as exogenous variables. We decided to try this direction by introducing Fourier terms as exogenous

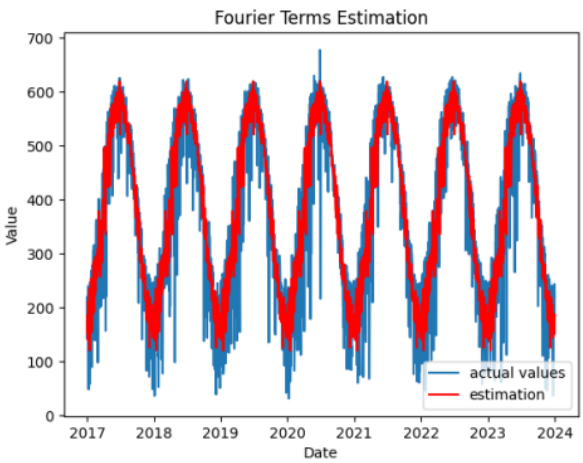
variables to the ARIMA model, replacing the seasonality in SARIMA with Fourier terms. We clarify that this subsection has nothing to do with the assignment given in Part 3, as the exogenous variables are derived from the behavior of the data, which is not motivated from any external sources.

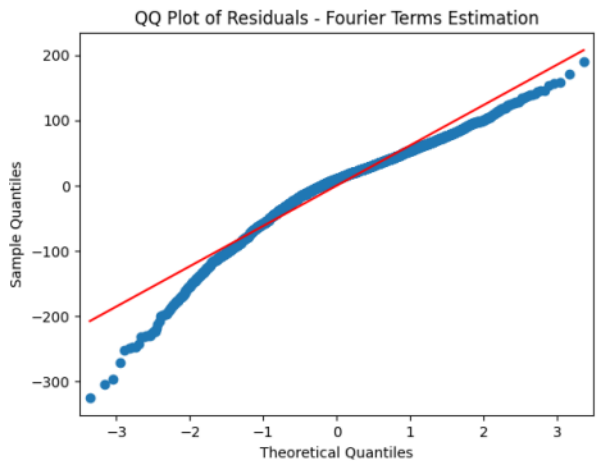
We first show the motivation behind the usage of fourier terms. In the analysis we performed, we got a high coefficient on the first cosine term when evaluating the coefficients of 183 terms.

Observing the results of the Fourier estimation, we see that the data is highly dependent on the leading Fourier terms. However, it seems that the residuals do not entirely follow a normal distribution. Consequently, we decided to take the first two terms of the Fourier analysis and use them as exogenous variables. The run times were a lot faster, and therefore a grid-search over the parameters was possible. Based on our observations in the ACF/PACF plots, we chose to start with ARIMA(1,0,1) with the exogenous variables of the first two Fourier terms with a yearly period, resulting in RMSE of **71.5**. After applying a grid search over AR,MA from 0 to 5, and number of terms in the set {2, 10, 50, 182, 365}, we found that the best RMSE was obtained by using ARIMA(4,0,5) with 2 Fourier terms. The RMSE achieved was 71.29. We also noticed that the AIC and BIC measures were not that different either (ARIMA(1,0,1): AIC=24165, BIC=24199. ARIMA(4,0,5): AIC = 24158, BIC = 24232).

Value	p-value	# of Term	Variable
381.21	0.00	0	Intercept
-204.50	0.00	1	cos
12.50	0.00	1	sin
4.73	0.01	8	sin
-3.75	0.04	9	cos
-4.29	0.02	18	cos
-4.54	0.02	25	cos
-3.98	0.03	26	cos
4.94	0.01	27	cos
-4.42	0.02	27	sin
-3.74	0.05	29	cos
4.23	0.02	33	cos
5.68	0.00	34	cos
4.74	0.01	35	cos
-5.29	0.00	40	cos
-5.00	0.01	41	sin
4.72	0.01	42	cos
-4.35	0.02	44	cos
-3.90	0.04	61	sin
4.57	0.01	66	sin
4.49	0.02	68	cos
3.78	0.04	78	sin
-3.67	0.05	79	cos
-4.38	0.02	96	sin

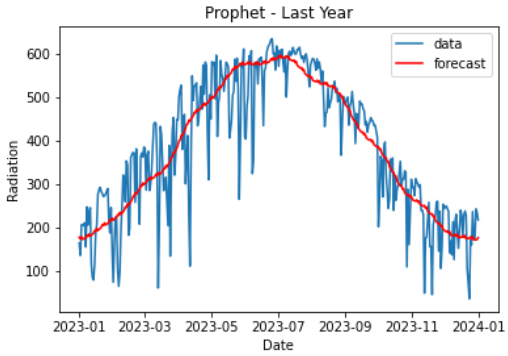
Note: Other attempts were made, to add the observation of last year as an exogenous variable, or the observation of 182 days prior as an exogenous variable. Alas, they did not prove to be effective, even when setting the starting parameters to give more emphasis to them. We believe that this is a direction which might be useful to explore in a more comprehensive manner, as this was not stated in the course.





Prophet

Another model we have learned about when dealing with seasonal data is the **Prophet** model by Facebook.



Manual Experiments over Hyper-parameters

First, we decided to try the default Prophet model. Comparing to SARIMA, the fitting procedure lasted merely seconds, making this model far more efficient. The run time was similar that of the ARIMA model. The prophet obtained **RMSE** of **71.85** on the test set.

We then tried achieving better results by adjusting the following parameters of the prophet model:

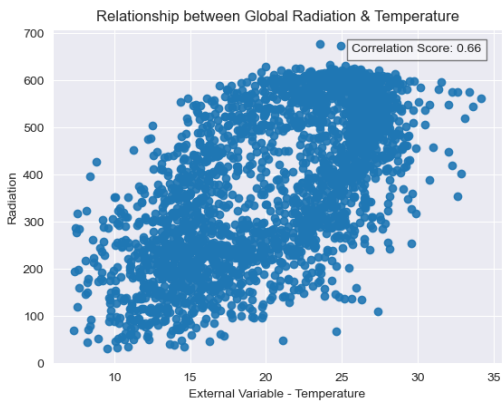
Period	$\in \{182, 365, 365.25, 720\}$
Fourier order	$\in \{2, 5, 20, 100, 180, 365\}$
changepoint prior scale	$\in \{0.01, 0.1, 1, 10, 100\}$
seasonality prior scale	$\in \{0.01, 0.1, 1, 10, 100\}$

We filtered out combinations based on their performance on the test test (which was the last 365 days of the dataset, i.e last year of documented data) All in all, the most optimal model on the test set was the Prophet model with the following hyper-parameters: period = 365, fourier order = 2, changepoint priorscale = 0.1 and seasonality prior scale = 0.1.

Incorporating an Exogenous Variable

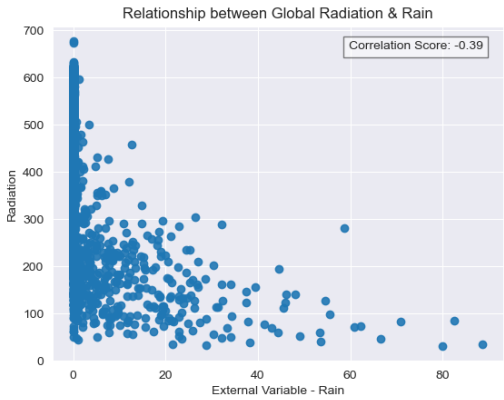
To incorporate an exogenous variable into the analysis, we checked 3 exogenous variables – **Rainfall**, **Temperature**, and **Sun duration**. The measures were available in the meteorological database and were all recorded at the Haifa Technion station, covering the years from 2017 to 2023. Much more variables were available in the database, and after researching for the most suitable we converged to check the most appropriate.

Temperature[2] affects atmospheric conditions and the density of air, which can influence the scattering and absorption of solar radiation. Higher temperatures generally lead to greater atmospheric instability and increased cloud formation, potentially reducing global radiation levels. Conversely, lower temperatures may result in clearer skies and higher global radiation. By including temperature as an exogenous variable, the model might capture the indirect impact of temperature on global radiation through its influence on atmospheric conditions.

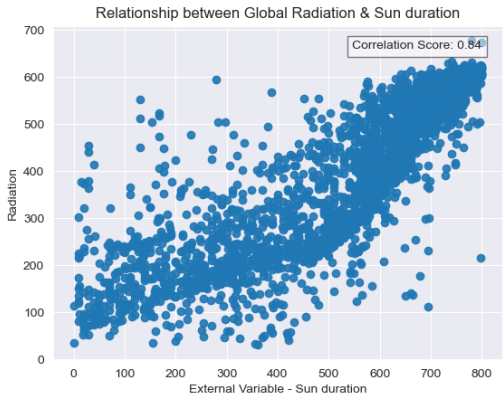


[2] Temperature day is defined from 18:00 UTC of the day previous the mentioned date, to the 6:00 UTC of the mentioned day.

Rainfall[3] is a crucial environmental factor affecting cloud cover and atmospheric conditions. High levels of rainfall often lead to increased cloud cover, which can block solar radiation from reaching the Earth's surface, consequently reducing global radiation levels. Therefore, incorporating rainfall as an exogenous variable allows the model to account for this inverse relationship between rainfall and global radiation, and might lead to better predictions.



Sun duration [4]represents the amount of time during the day that the sun is visible, which directly impacts the amount of solar radiation reaching the Earth's surface. The duration of the sun's caution refers to the measurement of the flux of direct radiation above a certain threshold [5] Reflects the duration of time, between sunrise and sunset, in which cloudiness, if present, does not interfere with the passage of direct radiation. Since there is more time for solar radiation to penetrate the atmosphere and reach the ground as the sun duration is longer, it is reasonable to check if the variable correlates with higher global radiation levels.



Statistical Methodologies and Explanations

Preprocessing

Preprocessing steps include converting dates to datetime format, calculating daily averages as aggregation, and handling missing values.

Aggregations

- **Temperature** – The original dataset includes 2 variables for temperature - chose to take the average between “min_termperature” and “max_temparture” measures from the original Temperature DataFrame. No missing values were detected.

Handling missing values

- **Rain** – Since the original measures didn’t include values for days without rain, we can safely say that missing data points are **MAR**, since missingness indicating no rain for the missing date. We manipulated the rain DataFrame so it will have all the dates between 1/1/2017 to 31/12/2023, with 0 values for the missing dates as appropriate controls.
- **Sun duration** – We couldn’t figure why specific dates were missing from the original DataFrame, so we assumed missingness of type **MCAR**. Since data wasn’t very

[3] Rainfall day is defined from 6:00 UTC of the mentioned day to the 6:00 UTC of the next day.
[4] Sun duration day is defined as the default notation for regular day (00:00-24:00).
[5] approximately 200 watts per square meter.

sufficient with size, we figure that imputation methods might increase the bias. We found a simple yet robust way to deal with the missingness – Linear spline interpolation. Linear interpolation is a suitable choice for completing missing data in time series datasets like sun duration. Its simplicity, based on the assumption of a linear relationship between consecutive data points, makes it computationally efficient and easy to implement. By preserving the overall trend and smoothness of the data, linear interpolation provides conservative estimates without introducing abrupt changes, ensuring that the interpolated values align closely with the underlying pattern of the time series. We chose to use Linear Spline interpolation due to its ability to capture more complex relationships between data points. Unlike linear interpolation, which assumes a straight line between consecutive points, spline interpolation fits piecewise linear segments between neighbouring data points. This flexibility allows spline interpolation to better accommodate nonlinear trends and fluctuations in the data.

Statistical Analysis and Modelling

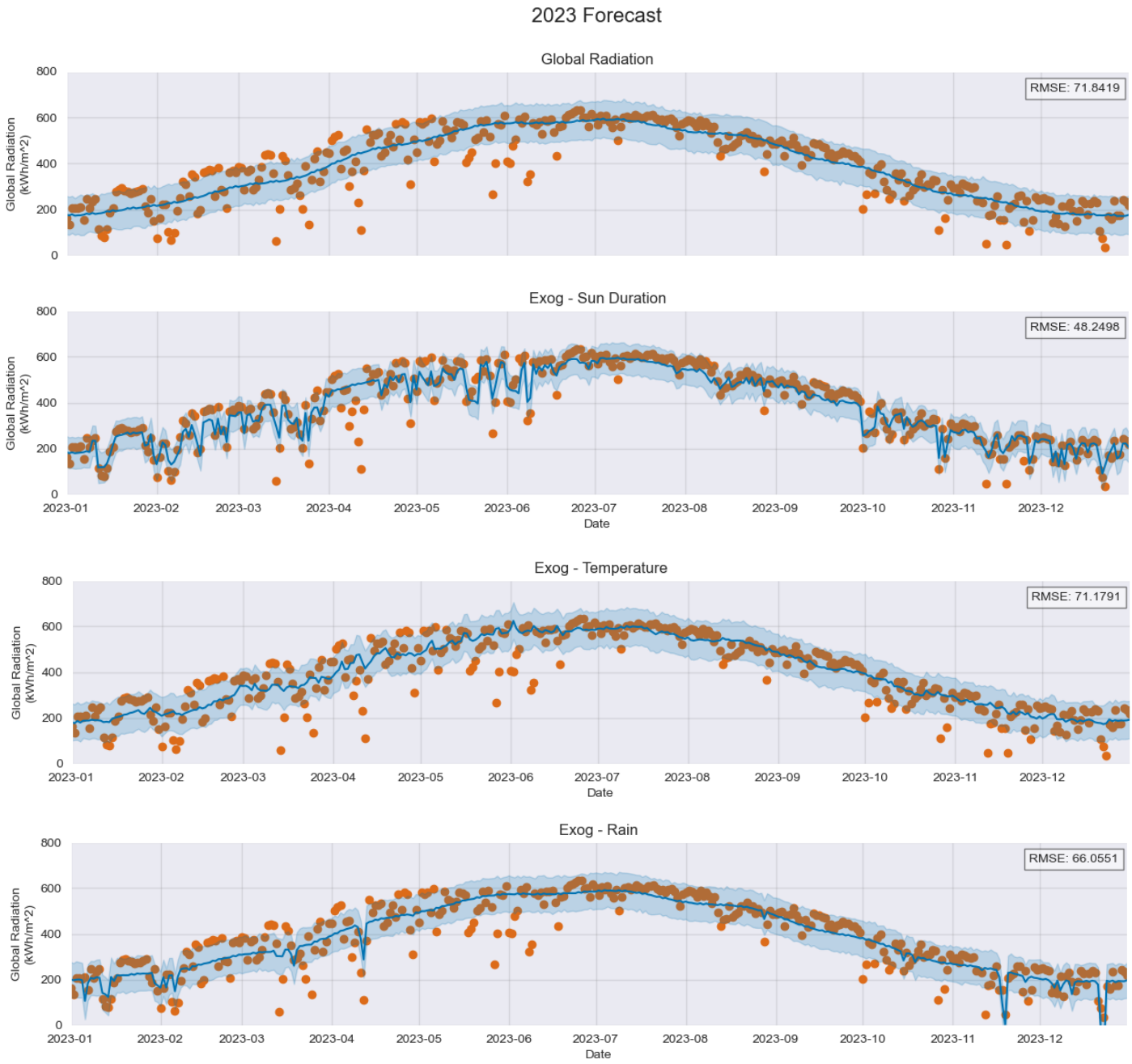
First, due to lack of time and computational limitations[6], only the Prophet model was taken under consideration. The model benchmark was the basic prophet model on the original time series.

Exogenous variables with a strong correlation with global radiation are more likely to improve the forecasting accuracy of the Prophet model. Their inclusion enables the model to better capture the complex relationships and dependencies present in the data, resulting in more accurate predictions. So, to determine the most suitable exogenous variable for inclusion in the Prophet model, we first calculate the correlation score between each exogenous variable (rainfall, sun duration, and temperature) and the target variable, global radiation. The correlation score helps assess the strength and direction of the relationship between each exogenous variable and global radiation. As demonstrated in the correlation plots above, the most correlated variable is Sun duration, with 0.84 correlation score.

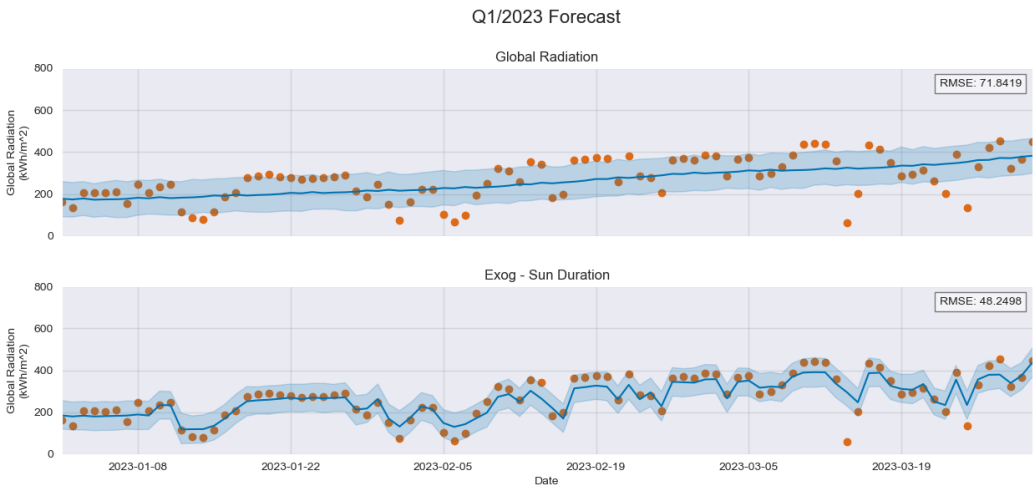
Next, we compare the performance of the Prophet model on the dataset in two scenarios: one without the exogenous variable and another with the exogenous variable included using the 'add_regressor' method of the Prophet model instance. For model evaluation, both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) functions were used.
Using both RMSE and MAE to evaluate time series predictions with Prophet provides a more comprehensive understanding of the model's performance. RMSE captures the average magnitude of errors and penalizing large errors more heavily, while MAE considers the average absolute difference between predictions and actual values, providing a balanced view of errors. Considering both metrics allow us to check model's performance from different perspectives.
The metrics are available in the python Skicit-Learn package.

Given that, we were able to quantify the forecasting accuracy of each model. Moreover, Exogenous variables that exhibit consistent trends or patterns over time are more likely to align closely with the fluctuations in global radiation. As a result, their influence on the forecasting performance of the Prophet model is reflected in both the correlation score and the mentioned metrics, leading to consistent findings between the metrics.

[6] More on that in “Results and Discussion” section



Based on the mentioned metrics, the exogenous variable which demonstrated the best contribution to the original model was Sun Duration.



Results and Discussion

Models Comparison

Summing up the performances of the models, we have SARIMA with RMSE score of 75 (approx.), and both ARIMA with Exogenous variables and Prophet with RMSE score of 71 (approx.) on the test set.

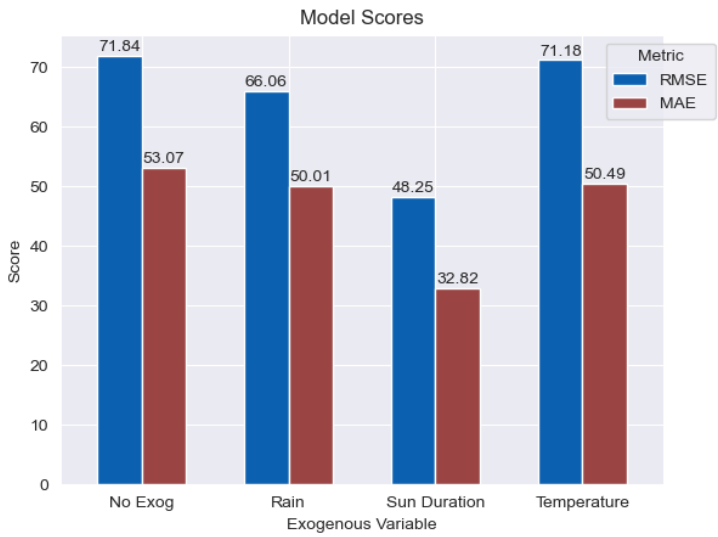
It appears that the addition of Fourier terms contributed to the performance of the models, as ARIMA with Fourier terms as exogenous variables outperformed the SARIMA model. It seems that the Prophet model performed quite similarly to the ARIMA model with the Fourier terms. This is likely since both try to capture the seasonality of the data with periodic components.

It is worth mentioning that hyper-turning the Prophet model did not provide any significant improvement in performance. This is to no surprise since, as stated in the beginning of this paper, sine and cosine waves are very descriptive tools when it comes to movement of celestial entities (i.e. the sun from which most of the documented radiation received comes from).

It is, however, important to state that the addition of Fourier terms reduced the ability of the model to rely on other things. We can see from the comparison of SARIMA to the other variation that were attempted ($ARIMA$ + Exogenous Variables, $Prophet$ (both default and hyper-tuned model) that SARIMA was more "daring", there were a lot more bumps and spikes, whereas Fourier oriented models tended to find a wave and stuck to it. Therefore, should a user require a model that imitates a noisy behaviour, removing the Fourier terms might be a viable option.

The exogenous variable tested only for the prophet model, due to computational limitations (memory usage exceeded) and for moderate simplification.

Using Sun Duration as exogenous variable yielded the lowest **RMSE** score, and lowest **MAE** score compared to the other variables. The metrics score indicates that the model with Sun Duration as an exogenous variable better captures the underlying patterns and variability in global radiation levels, making it the most appropriate variable to include in the forecasting model. Combining the initial motivation for Sun duration together with Prophet model scores and the correlation score (with the original global radiation series), we can suggest the variable as an appropriate exogenous variable, compared to the other mentioned variables.



Conclusions

In this paper, we have attempted to predict the average daily radiation which was measured in the Technion institution by applying time series analysis methods we were presented in the course. We have used two traditional models and several of their variations, and in addition, tried to add some data-driven seasonal modeling by using Fourier terms (which was made viable due to the behavior of the data).

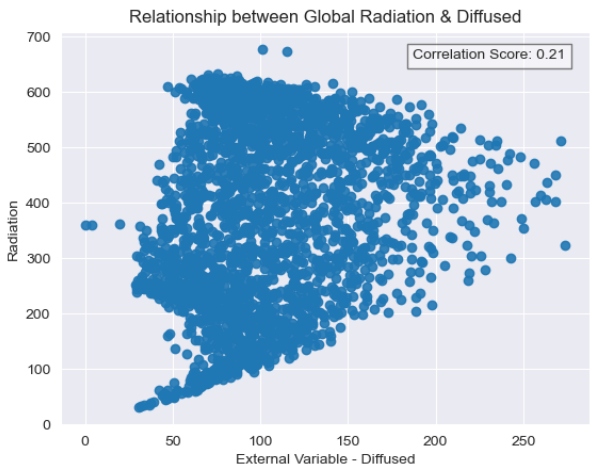
Moreover, several variables were examined as exogenous variables to one of the models (Prophet), with one of them, which proved to be the most contributing to the predictions, reducing the RMSE by a significant amount of **30**.

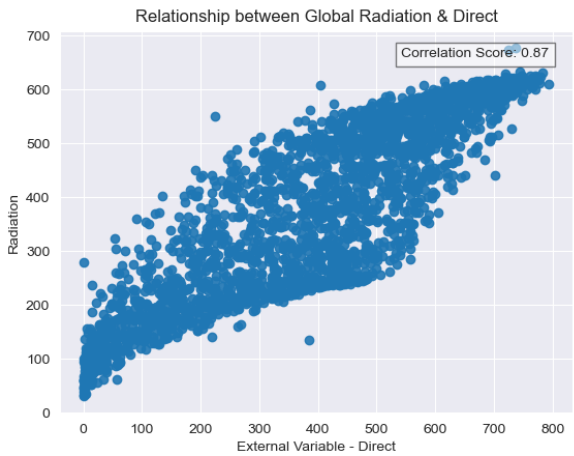
References

1. Israel Meteorological Service databases
2. Detailed guidebook to the user of IMS data (Hebrew)
3. About the radiation database (Hebrew)
4. Israel Meteorological Service Database API
5. Linear Interpolation & Spline linear interpolation for missing data
6. Spline linear interpolation
7. Mean squared error regression loss - Skicit-Learn Documentation
8. Mean Absolute Error regression loss – Skicit-Learn Documentation

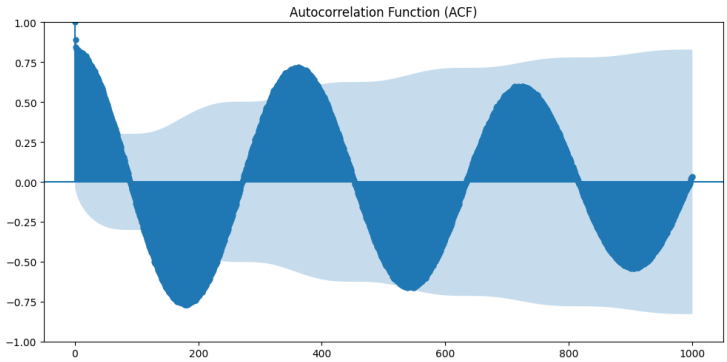
Appendix

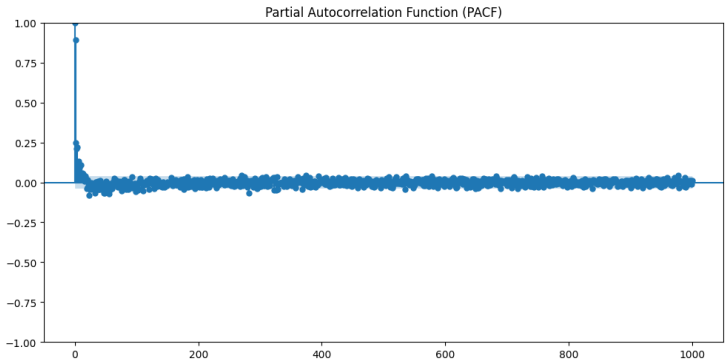
1. Correlation of Global Radiation with each of its subcomponents – Direct and Diffused:





2. Global radiation Autocorrelation plot and Partial Autocorrelation plot
(lag = 1000)





3. Global Radiation box plot by Quarters: