Data Analysis and Green Computing: Profiling HPC Power and Tracking CO2 Emissions

Green Team Report

Supervisor: Mohsen Seyedkazemi Ardebili

April 2024

Contents

Abstract

In the following document we try to collect all possible observations and intuitions related to our work, whose goal is to profile High Power Computing power consumptions and track CO2 emissions related to it.

Contributors:
Devletiyarov Baibek
Liberatoscioli Martina
Squarzoni Lorenzo
Zavaroni Sofia

1 Tools and methods used for the analysis

To start, a brief introduction to the tools and methods employed for data analysis is provided.

1.1 STL (Seasonal-Trend Decomposition)

STL (Seasonal-Trend Decomposition) is a technique used in time series analysis to break down a dataset into three main components: trend, seasonality, and remainder.
1. **Trend**: This component captures the long-term direction or progression of the data, indicating whether it is increasing, decreasing, or staying constant over time.
2. **Seasonality**: Seasonality refers to the periodic patterns or fluctuations that occur at regular intervals within the data, such as daily, weekly, or yearly cycles.
3. **Remainder**: The remainder component represents the variability in the data that cannot be explained by the trend or seasonality. It captures the random fluctuations or noise present in the data.
By decomposing the time series data into these components, STL allows to understand the underlying patterns and structures, making it easier to analyze and model the data effectively.

1.2 Meta's Prophet

Prophet is a forecasting tool developed by Facebook, designed to handle time series data with seasonal patterns and uncertainty. Its key points are:
1. **Automatic Seasonality Detection**: Prophet automatically detects seasonal patterns in the data, making it suitable for datasets with irregular or changing seasonalities.
2. **Trend Flexibility**: Users can specify various components of the time series, including holidays and special events, which are incorporated into the forecasting model.
3. **Uncertainty Estimation**: Prophet provides uncertainty intervals around the forecasted values, helping users understand the range of possible outcomes.
4. **Scalability**: It's designed to be scalable and can efficiently handle large datasets.
In summary, Prophet simplifies the time series forecasting process with an intuitive interface and powerful forecasting capabilities, making it suitable for both beginners and experienced analysts.

Prophet automates seasonal modeling for forecasting, while STL offers manual control for detailed trend analysis. Choose based on data complexity and desired level of control.

2 Power and Energy Analysis

The data distribution spans from April 2020 to October 2022, showing a discrete amount of information that can be enough to make some initial observations and guessings. While there isn't a discernible pattern across the entire period, comparing plots for specific nodes (r205n01), the sum of all nodes, and individual racks reveals certain trends.
The lines in the plots that have a strange or unusual pattern are the result of the horizontal and vertical interpolation applied to fill some empty spaces in the distribution. No further manipulations have been applied to the data.

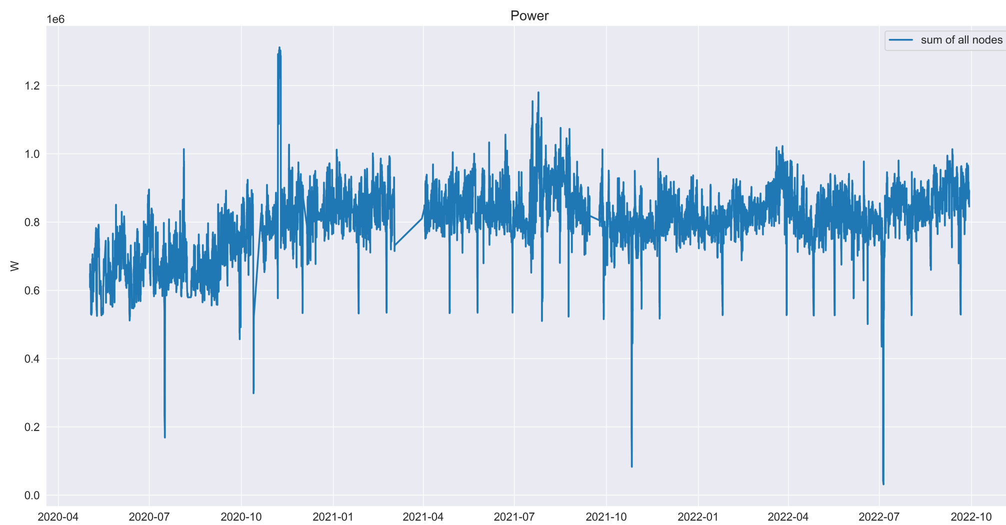

Figure 1: PWR total value (sum of all nodes in the server)

count 2.110200e+04
mean 7.997859e+05
std 1.008300e+05
min 3.101649e+04
25% 7.600741e+05
50% 8.096636e+05
75% 8.581640e+05
max 1.311606e+06

In terms of pure power consumption, we see a visible peak that reaches a value of 1.3 MW, while the general mean stays close to 0.8 MW. The data fluctuates a lot throughout the days, but it is difficult to find any particular pattern or repetition at this level of depth; what we can guess is that all or most of the lowest points in the plot are given by a moment or period of maintenance for the server, while the highest values might indicate an episode of testing for the capabilities of the server in terms of maximum computational power.

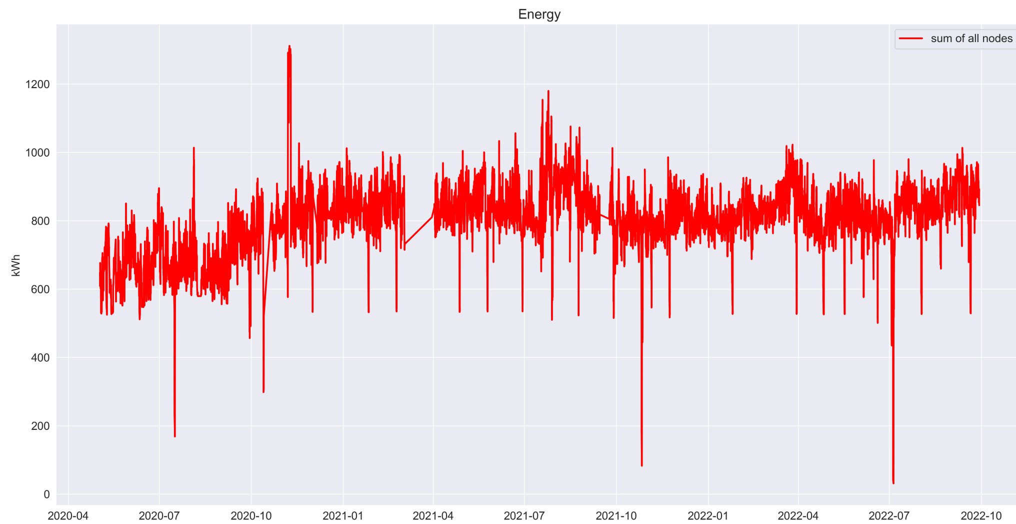

Figure 2: E total value (sum of all nodes in the server)

count 21102.000000
mean 799.785924
std 100.830046
min 31.016491
25% 760.074139
50% 809.663640
75% 858.164028
max 1311.606144

The energy computation was conducted based on power consumption data to derive energy values in kWh.
Distinct regions in the plot exhibit consistent energy levels around the mean, notably in June and August, alongside regions with negative peaks (e.g., July and October) and others with positive peaks. Understanding these peaks in the context of high-power computing offers valuable insights. For example:

- Cooling needs vary with external conditions like temperature;
- Peaks may coincide with hotter periods, necessitating more energy for cooling.
- Changes in online service or data processing demand can influence energy consumption, e.g., heightened activity leading to increased power usage.

The absence of a pattern in a dataset can be scientifically attributed to various factors including randomness, complexity, noise, limited sample size, underlying dynamics, data quality issues, and confounding factors. These factors can obscure any underlying trends or patterns in the data, making it challenging to discern meaningful insights.

2.1 PWR r205

At both rack and node levels in datacenters, the power consumption fluctuates around a baseline representing the average usage.
This is due to varying workloads, with peaks corresponding to high-demand periods and valleys to low-demand one, to transient events and efficiency optimization strategies.

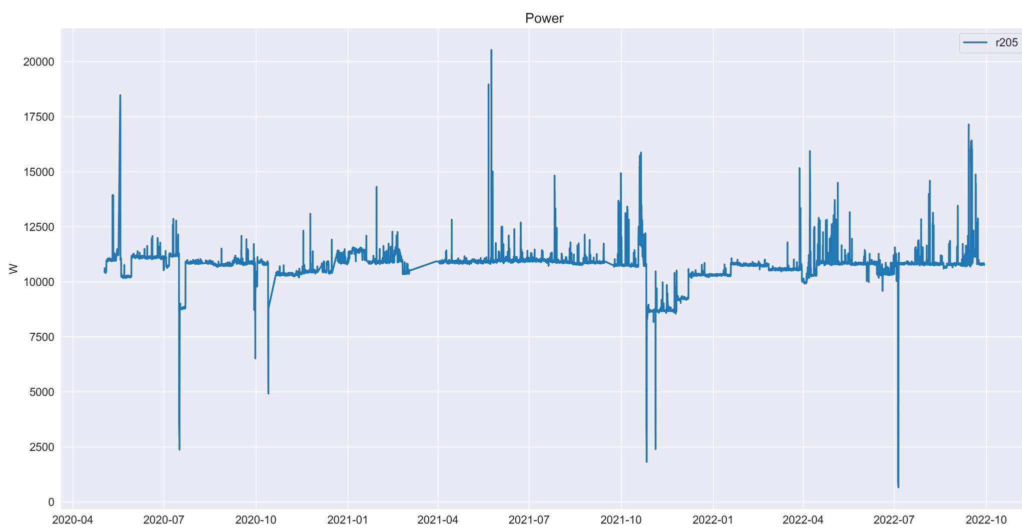

Figure 3: PWR r205

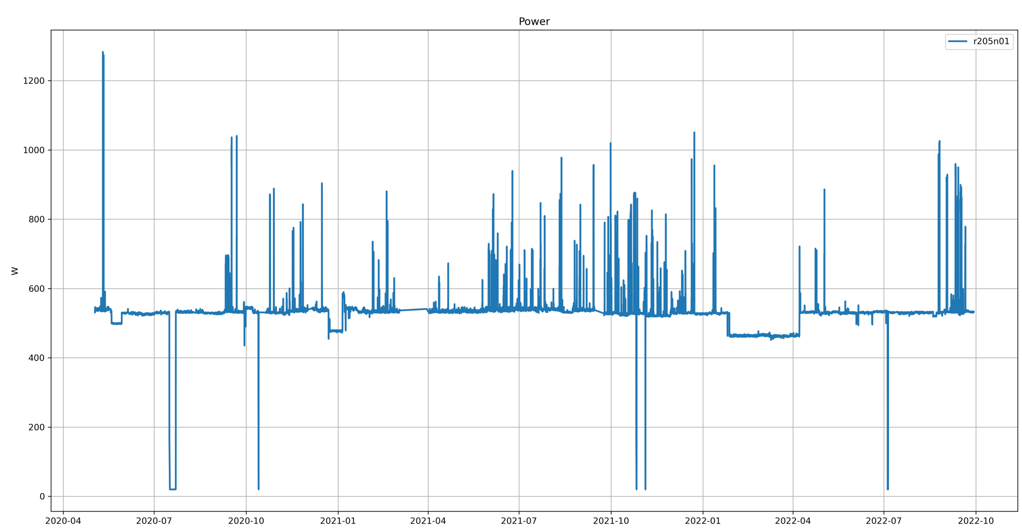

Figure 4: PWR r205n01

2.2 PWR r206

Comparing the plots of the first rack (r205) with respect to the others it seems to be that the only one with a regular and stable power consumption is the r205.
Looking at a different rack as the r206 in figure 5 (but every other rack is more similar to this) there is a much different and oscillating plot.

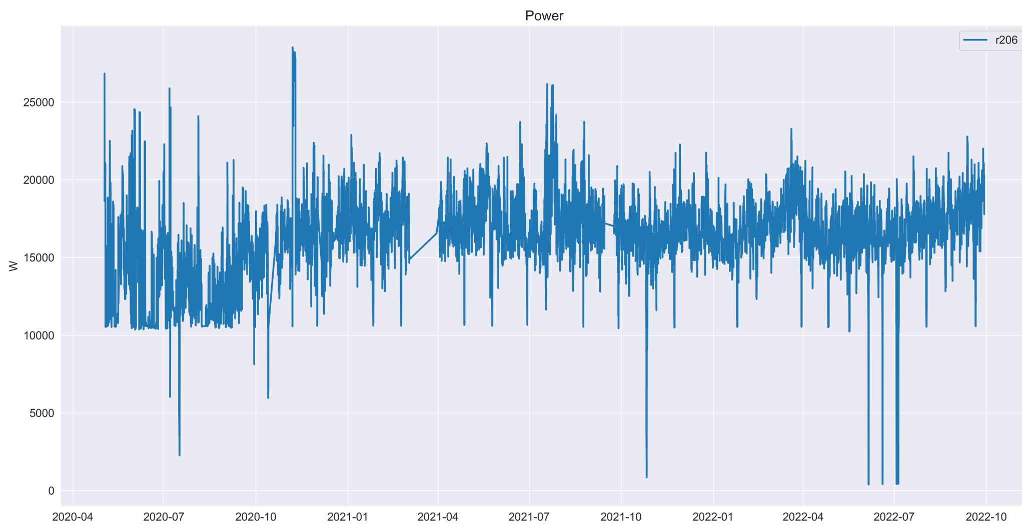

Figure 5: PWR r206

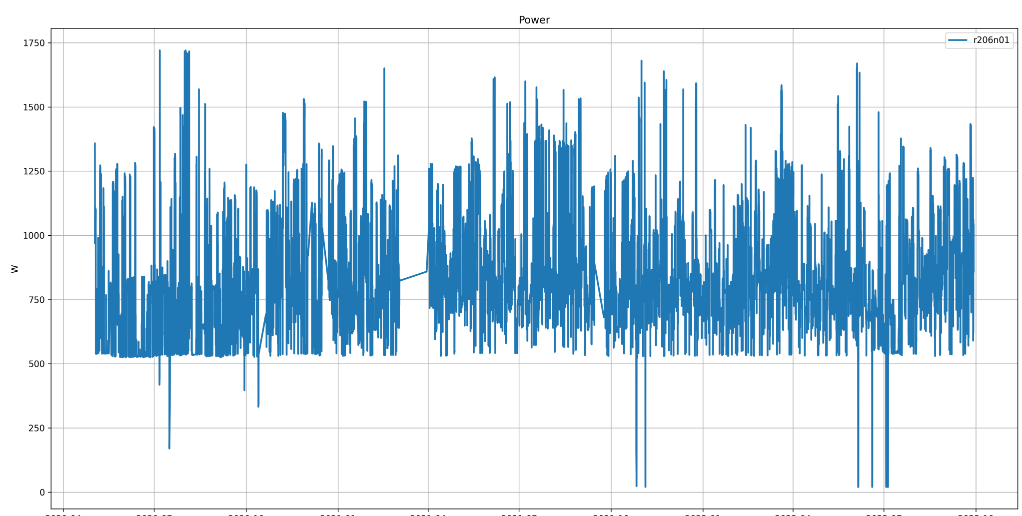

Figure 6: PWR r206n01

2.2.1 Comparison r205 and other racks

The first rack in a data center often exhibits consistent power consumption due to several factors:

- **Login nodes**: it is possible that the first rack contains login nodes. Due to this it exhibits a more regular pattern in power consumption since login nodes are primarly used for user access, job submission and management tasks which tend to have a more predictable load than the highly variable computational tasks running on compute nodes. In addition the workloads on login nodes are generally less intensive and more uniform over time, leading to more consistent power usage patterns and the hardware in login nodes is often optimized for efficiency and stability, which can contribute to a more predictable power consumption profile;
- **Critical Infrastructure Equipment**: the first rack often contains critical equipment for infrastructure management, such as main routers or switches, which require a constant and reliable power supply. In addition data centers are designed with power distribution systems that ensure the first racks receive stable power and are not affected by load variations that may occur further down the distribution chain;
- **Thermal Management**: The first racks are strategically positioned to benefit from optimal airflow, maintaining a constant temperature and reducing the risk of overheating. Consistent thermal conditions contribute to greater power regularity, as fluctuations in temperature can impact the energy efficiency of equipment;
- **Facility Design**: The placement of racks and airflow management within the data center play a role in maintaining power consistency. Proper facility design ensures efficient airflow, contributing to stable power consumption;
- **Uninterruptible Power Supplies (UPS) and Backup Generators**: Data centers employ UPS and backup generators to ensure uninterrupted and regular power supply to racks, safeguarding sensitive equipment from outages or voltage spikes;
- **Rack Cabinet Design**: The choice and arrangement of rack cabinets can also influence power regularity. Enclosed racks, for instance, support better cooling efficiency and power distribution compared to open-frame racks, contributing to consistent power consumption.

All these factors work together to ensure that data centers operate with consistent power, which is essential for the reliability and performance of the services they provide.

2.3 PWR r206n01 STL

Even making a seasonal-trend decomposition it's hard to highlight any specific trend, since the data is full of variations and outliers. Taking advantage of a different tool we'll try to make any seasonality clearer.

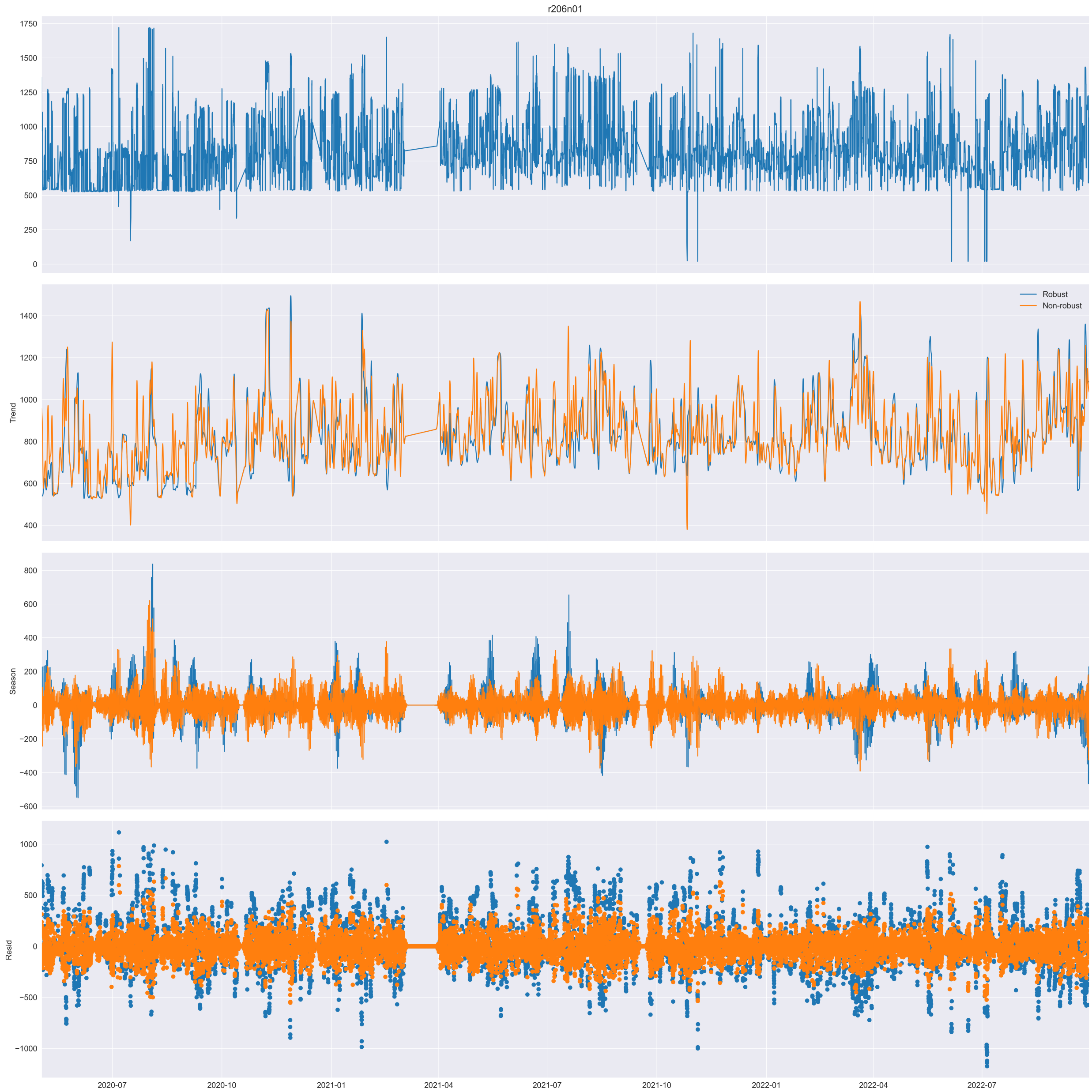

Figure 7: PWR r206n01 STL

2.4 PWR analysis using Meta's Prophet

Thanks to the usage of this method we are able to extract a clearer representation of the possible trends of power consumption in our server: good choices, purely looking at the following data, might be to increase the nodes' usage during the night, during week days and also during hot seasons, all periods in which power consumption results to be lower.

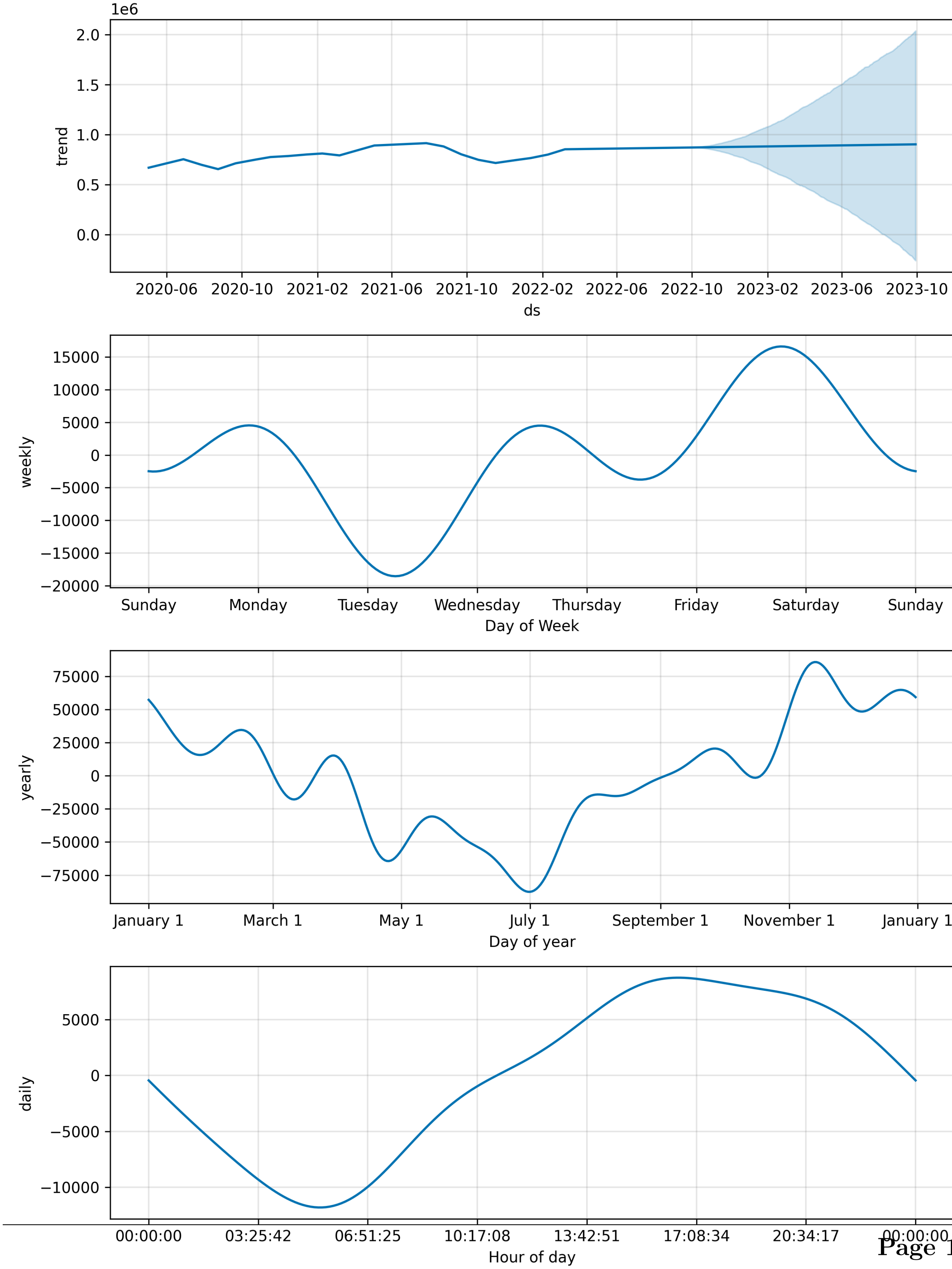

Figure 8: PWR trends

Analyzing these trends aids in optimizing energy management strategies and mitigating environmental impact in high-power computing environments. For example implementing energy optimization algorithms could reduce consumption during certain periods, resulting in negative peaks; while aiming to a better server energy efficiency can lead to improvements in server and cooling efficiency and alter the graph's shape over time, potentially reducing energy consumption peaks.

3 Carbon Intensity Analysis

The data distribution goes from 2021-01 to 2024-01.
In this case data hasn't been manipulated in any way, as no interpolation has been applied. With carbon intensity is much easier to highlight patterns and trends; in this case we are considering the carbon intensity related to direct power and energy usage, specific for our country (Italy) and area (Emilia-Romagna).

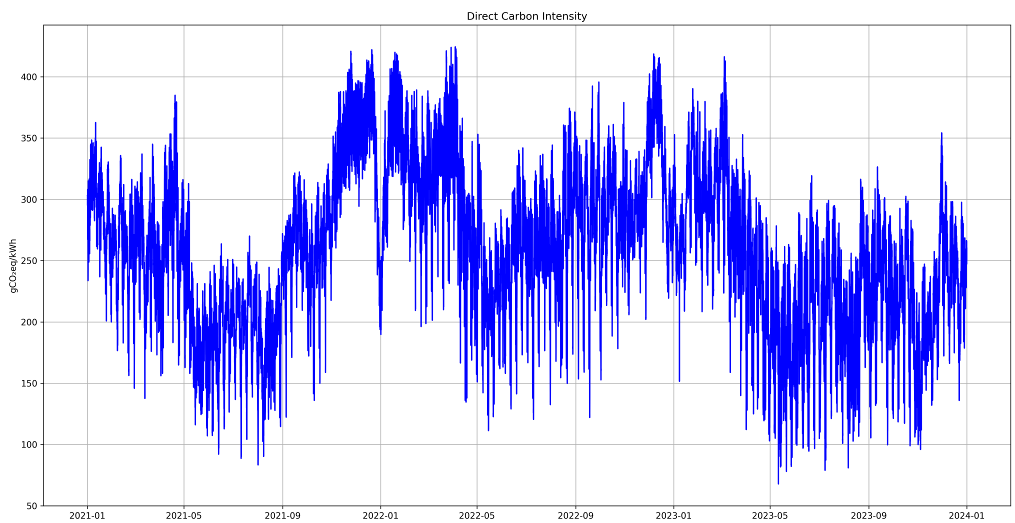

Figure 9: CI direct

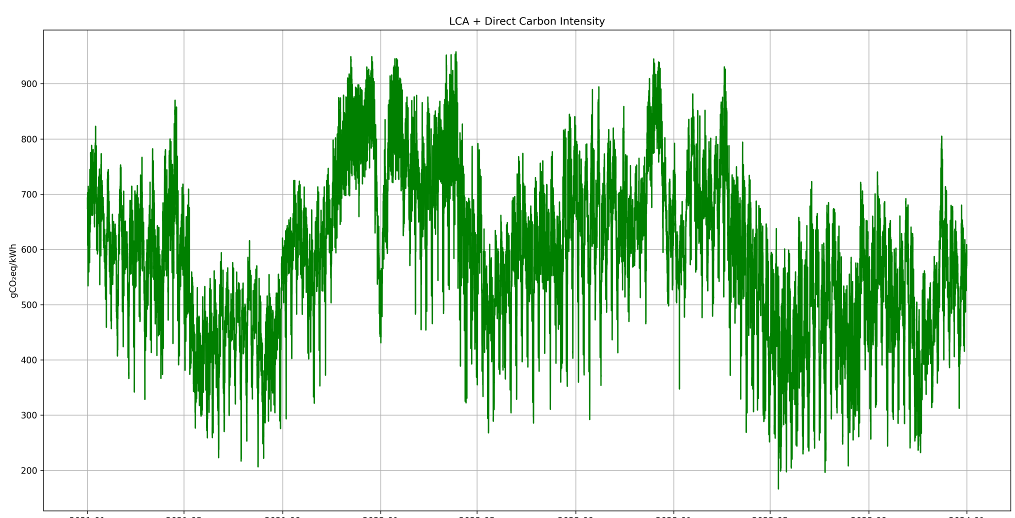

Figure 10: CI LCA + direct

count 26280.000000
mean 258.794563
std 63.450587
min 67.790000
25% 214.945000
50% 260.075000
75% 302.832500
max 424.440000

3.1 CI comparison

Both direct and indirect (LCA) carbon intensity share a similar pattern, showing equal highs and lows in time. Also the sum of the two distributions has the same behaviour.

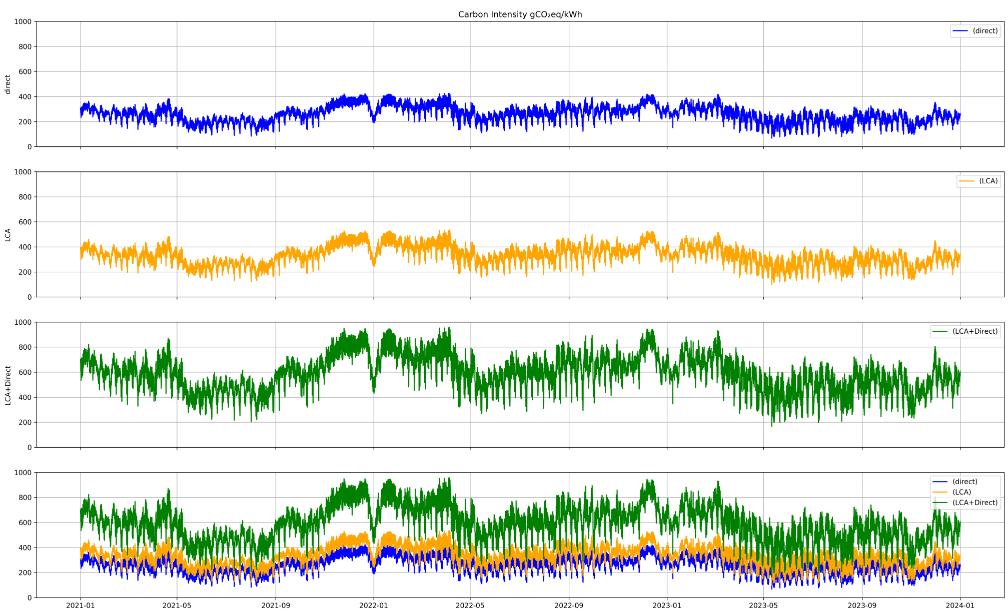

Figure 11: CI comparison

3.2 CI STL

Taking advantage of this decomposition we can easily guess the following assumptions: carbon intensity has a strong seasonality, showing a visible decrease during the warmer months and the opposite during cold ones; in addition there seems to be a strong decrease in the CI every January.

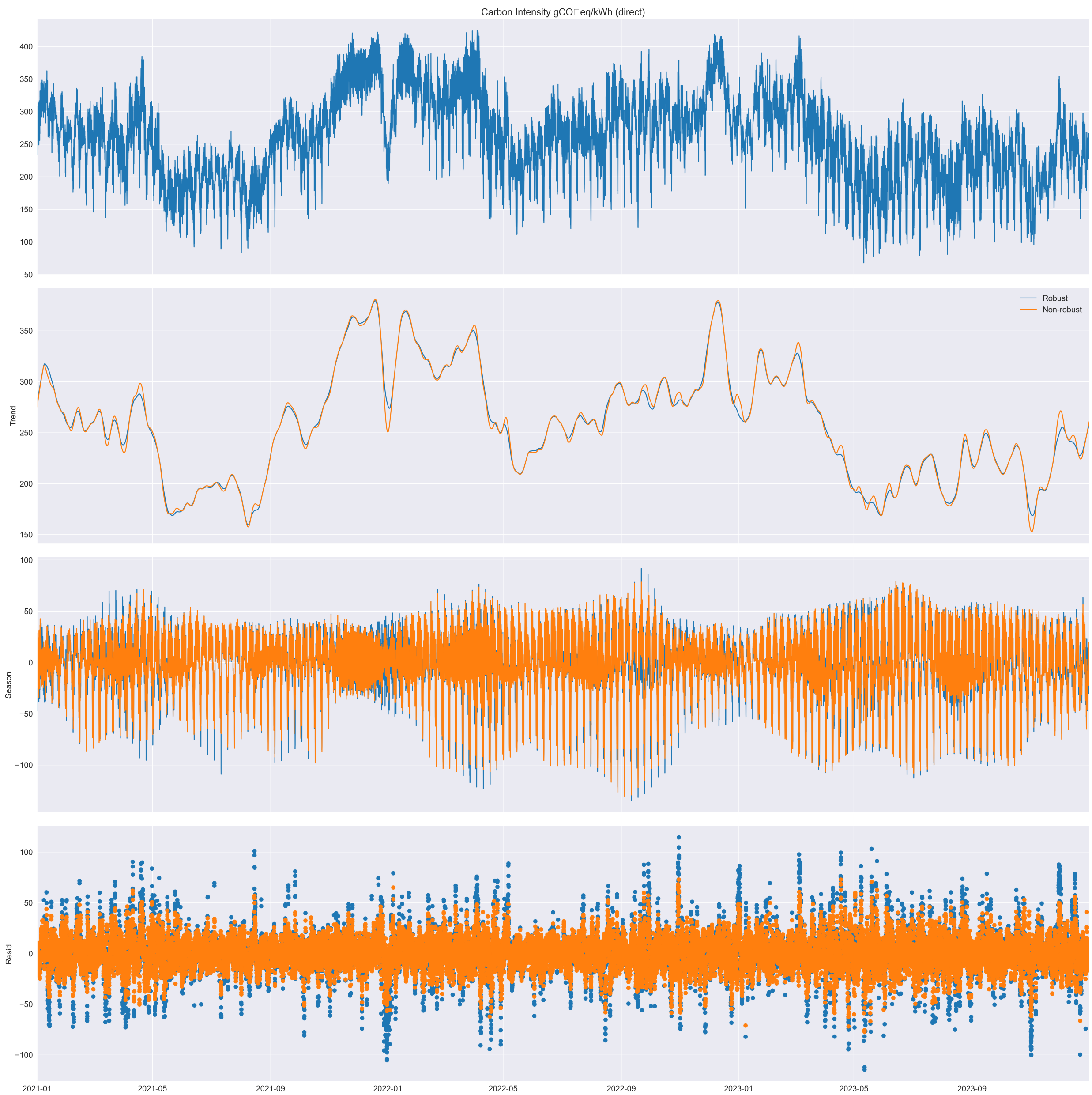

Figure 12: CI STL

3.3 CI analysis using Meta's Prophet

At deeper level (daily, weekly or monthly), much clearer patterns are visible:
During week days the carbon intensity is higher with respect to weekends, as well as the CI

contribution decreases during the lunch time (noon). Finally we can better visualise what we guessed before: during the year we see a couple of peaks, near February and December, with a lowering during January, contrary to what happens from May to September, where the general CI contribution goes down.

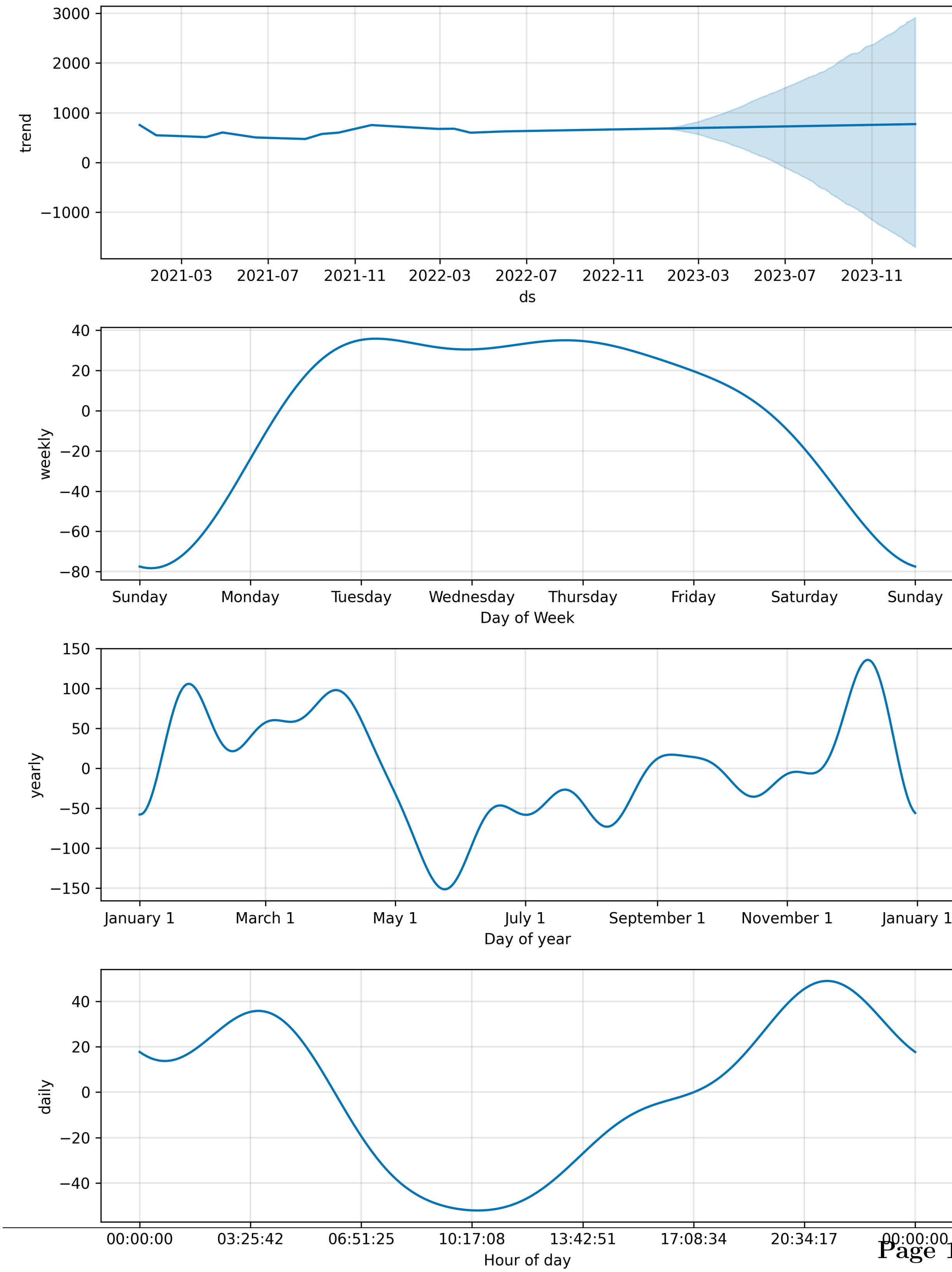

Figure 13: CI trends

4 Operational Carbon Footprint Analysis

Operation Carbon Footprint is the most important metric for the goal of our observation, since it is the result of combining energy usage (related to power consumption) and carbon intensity.
For this reason we will see a combination of the observations we did related both to power and CI, and our final suggestions will be a cross-reference of the previous data.

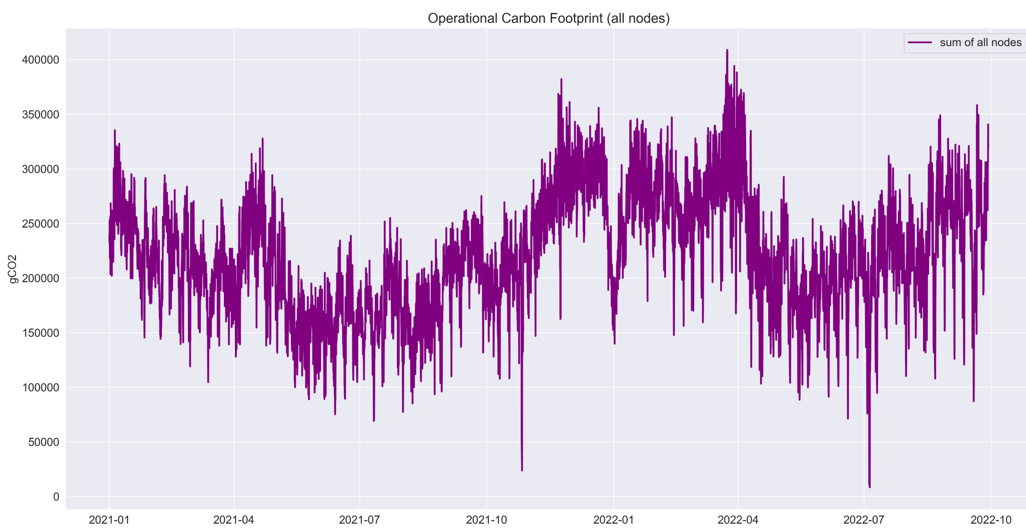

Figure 14: Cop total value (sum of all nodes in the server)

count 15263.000000
mean 220186.340067
std 54400.325364
min 8292.258962
25% 181522.725893
50% 218140.616056
75% 258846.427357
max 408868.699240

4.1 Cop r206n01 STL

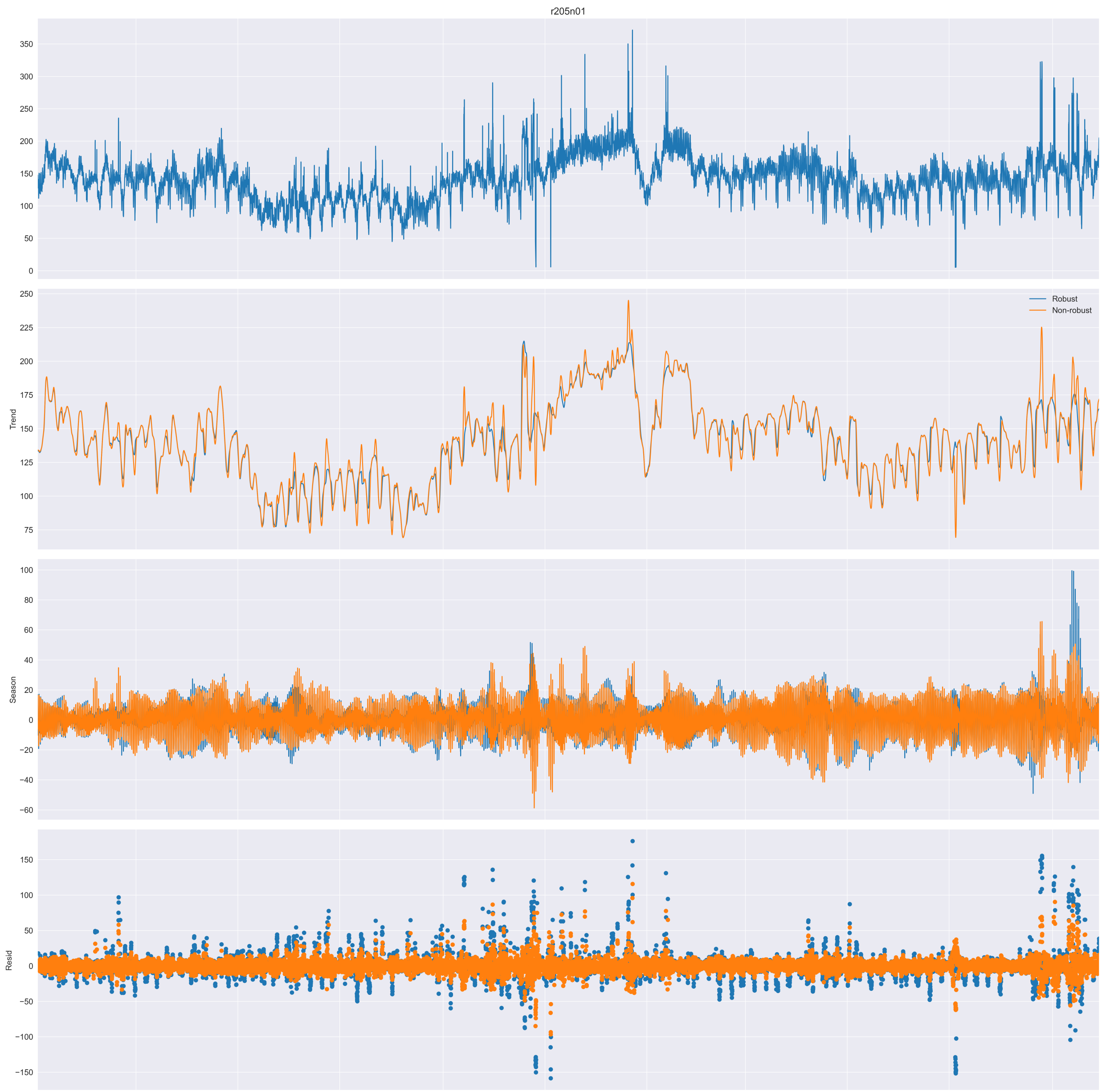

Figure 15: Cop r206n01 STL

4.2 Cop analysis using Meta's Prophet

As before, Prophet allows to better visualize the seasonality of our distribution:
in general the carbon footprint follows very much the carbon intensity distribution, showing the higher impact it has on the final pollution factor; looking at the week level plot, we clearly see a big lowering in equivalent CO2 production during sunday, while the same happens every day close to noon (same observations we did on the CI). At season and month level we confirm that the footprint lowers during summer.

Figure 16: Cop trends