Energy Analysis

As part of the interview process
For the role of Data Scientist at Habitat Energy

Jonathan Schuba
jschuba@gmail.com

Initial Thoughts and Observations

```
energy_data.describe()
Out[3]:
                SIP           DAM
count  89918.000000  89947.000000
mean      44.300704     44.946502
std       42.136848     24.044061
min     -153.890000    -38.800000
25%       27.170000     33.825000
50%       39.590000     42.530000
75%       55.600000     52.900000
max     4000.000000   1500.000000
```

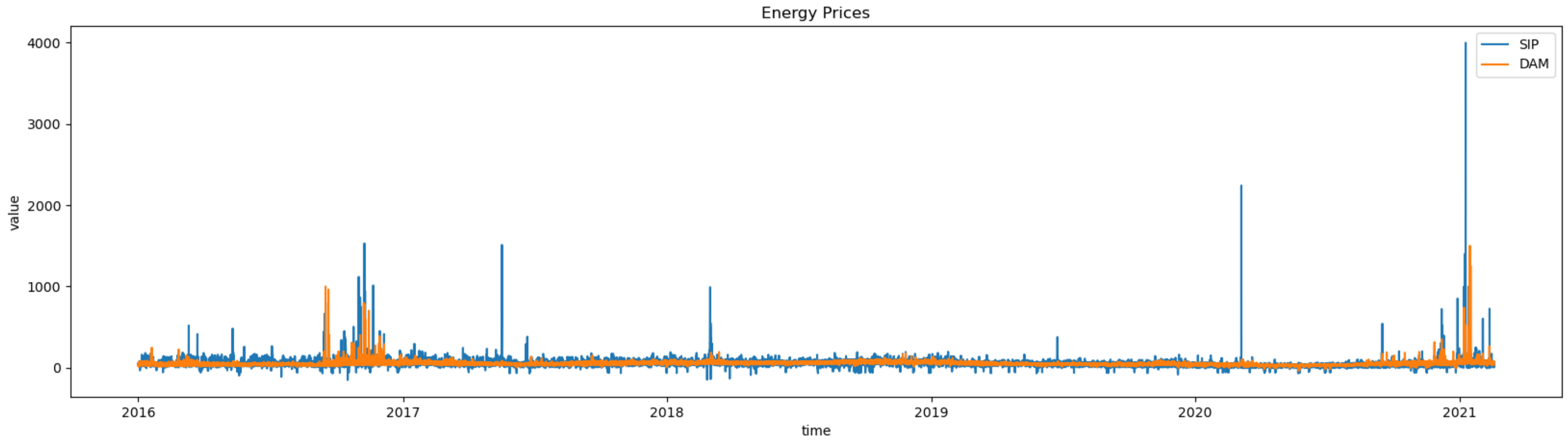
- The data set consists of 89953 timeseries datapoints, only a handful are Null.
- They cover 2016-01-01 to 2021-02-17, with a 30 minute step interval.
- It is not clear what the units are!
  - I will assume the units are £ / MWh, after some googling.
- Energy prices can go negative, meaning that someone will pay you to use power.
  - That makes sense, since some power plants (e.g., nuclear) are hard to throttle, and so supply may exceed demand at times.

After a little research:

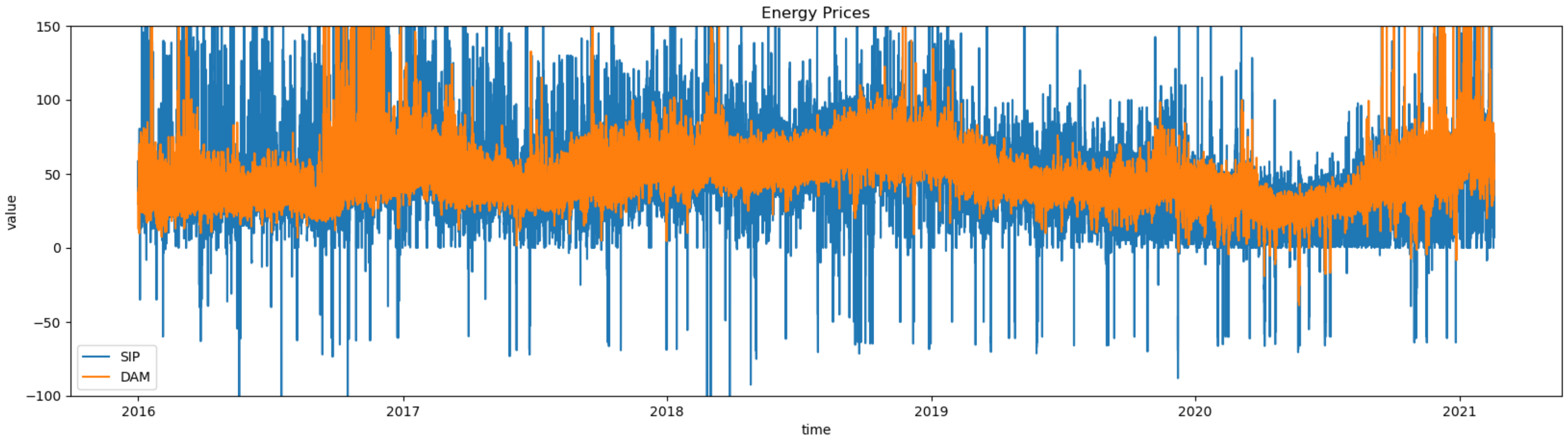
- DAM (Day Ahead Market) represents the contract price to buy or sell energy on the following day.
  - It is not clear if the datetime in the dataset is the time the contract is bought/sold or the time that the energy is to be delivered the next day.
  - I will assume the DAM datetime is the "next-day" time, similar to stock options pricing, where the date used is the expiration date.
- SIP (System Imbalance Price) is a mechanism to settle the difference between a party's contractual obligation and their actual volume. It appears to be a real-time calculation by the System Operator.
  - https://www.ceer.eu/documents/104400/-/-/66369fc0-516c-7b67-7106-0fa6e12c0511
  - I think it's more complicated than this, but I'm just going to treat SIP as the "real-time price" that would incentivize idle energy producers to come online and sell to the market. (And a negative SIP price would incentivize idle consumers to buy electricity.) Given more time, I would gather a more nuanced understanding.

  - Question: how does the regulator handle capacitance/inductance imbalance on the grid?

Initial Plots of the entire dataset

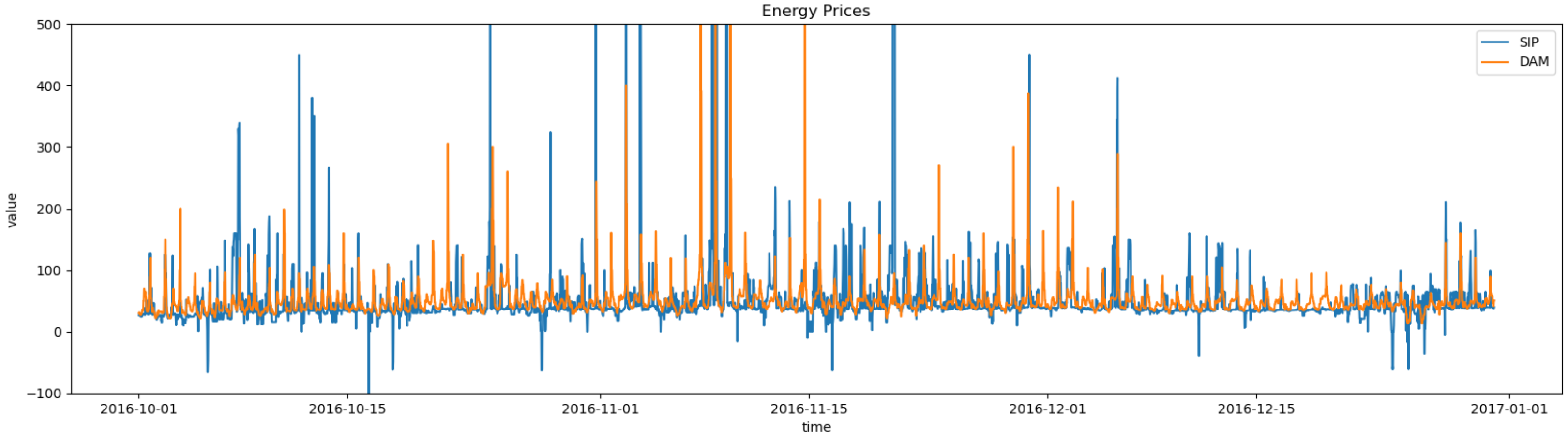


- There are a few large spikes in the prices.
- The spikes in the SIP price are larger, which makes sense since it's real-time.
- There are a few periods of great turmoil in the price:
  - October 2016 – December 2016
  - Early 2021



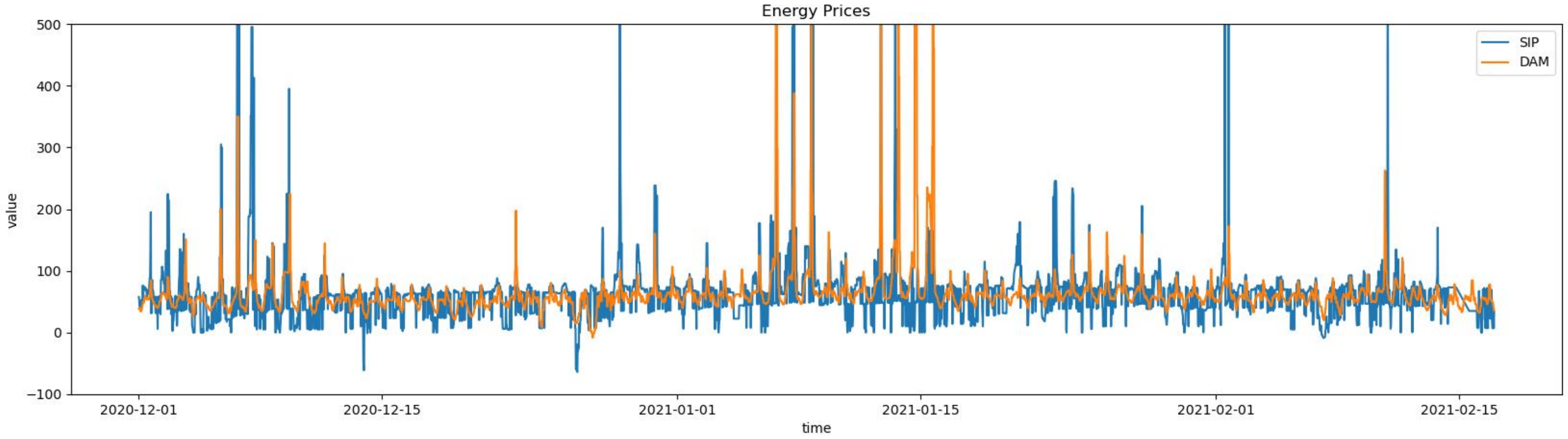
- Zooming in on the y-axis, we see that prices are generally between 30 and 70 *units*.
- The SIP price shows greater fluctuation, compared to the DAM price.
- I expected to see yearly periodicity, but it's not easy to spot here.

Zooming in on Periods of Turmoil



October 2016 to December 2016

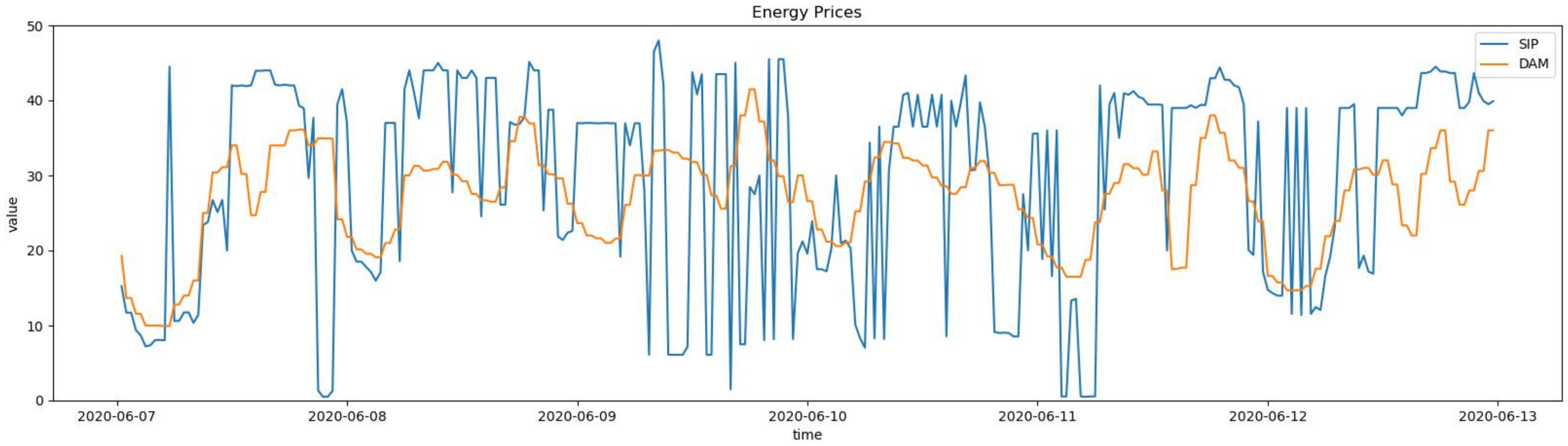
- We can see a daily periodic pattern on the DAM data. The SIP data shows some daily patterns, but is more erratic.
- The high price spikes occur at the same time of day but are higher on some days vs others.
- When negative spikes occur, they generally are half-a-day off phase from the high spikes.
- Without research, I don't know enough about the UK energy market to speculate on the causes for turmoil during this period. I do know that the Brexit vote was in June 2016, and this is probably a source of underlying uncertainty in the market.



December 2020 to End of dataset

- We can still see the daily periodicity.
- There are fewer spikes, but they are higher (exceeding the y-limit on the graph).
- Again, I don't know for sure what was happening around this time, but perhaps the spread of the COVID Alpha variant was impacting energy supplies. https://en.wikipedia.org/wiki/SARS-CoV-2_Alpha_variant#Spread_in_UK

Let’s look at a week



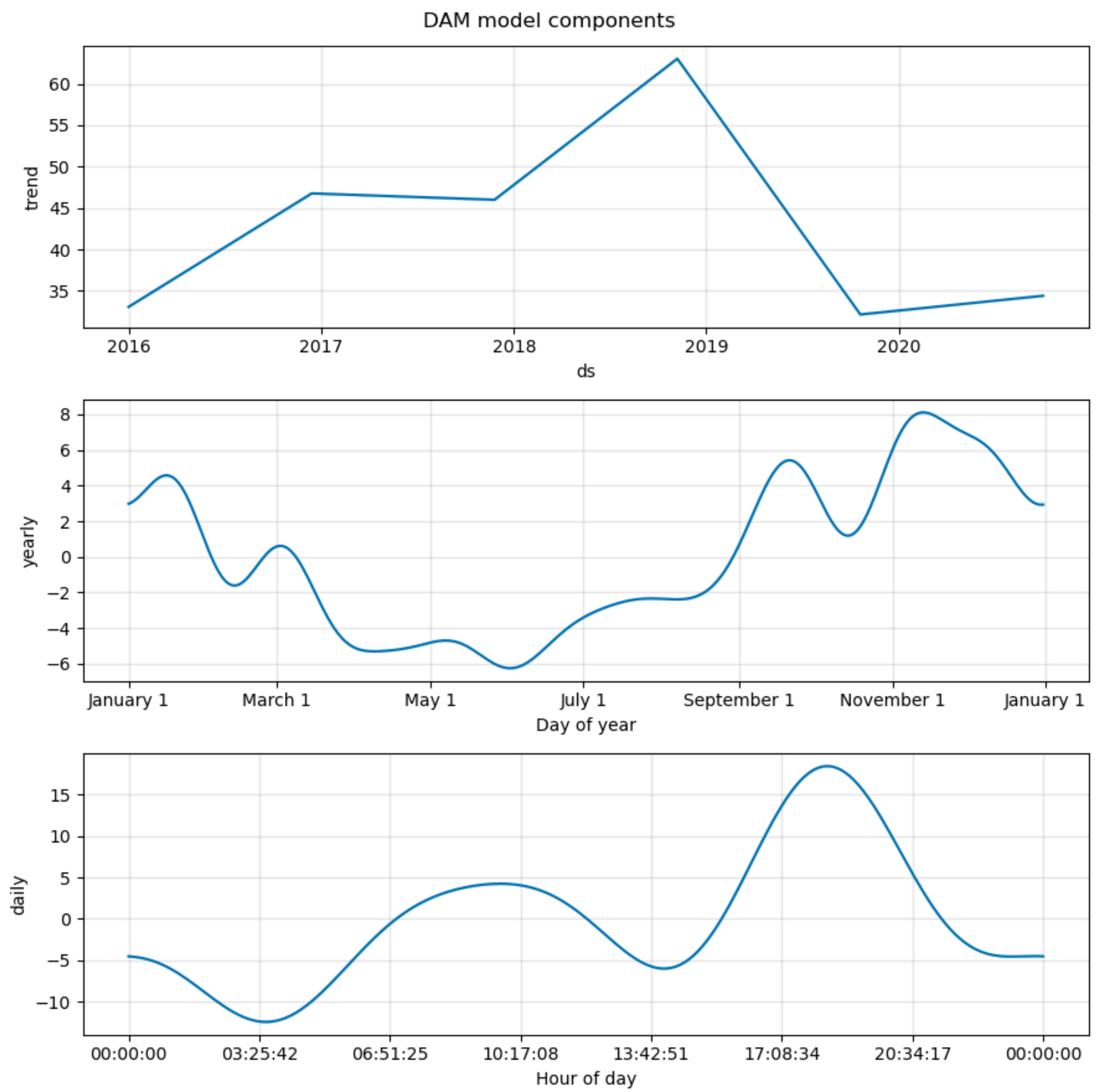
A typical week in June

- The daily periodicity is apparent.
    - Prices rise in the morning, as people wake up and use power.
    - Price falls in the afternoon, before peaking again in the evening.
- The SIP price spikes up and down, around the DAM price. Presumably, the regulator must incentivize or disincentivize additional capacity to adjust for periods when demand is different from forecast, or when suppliers fall short.

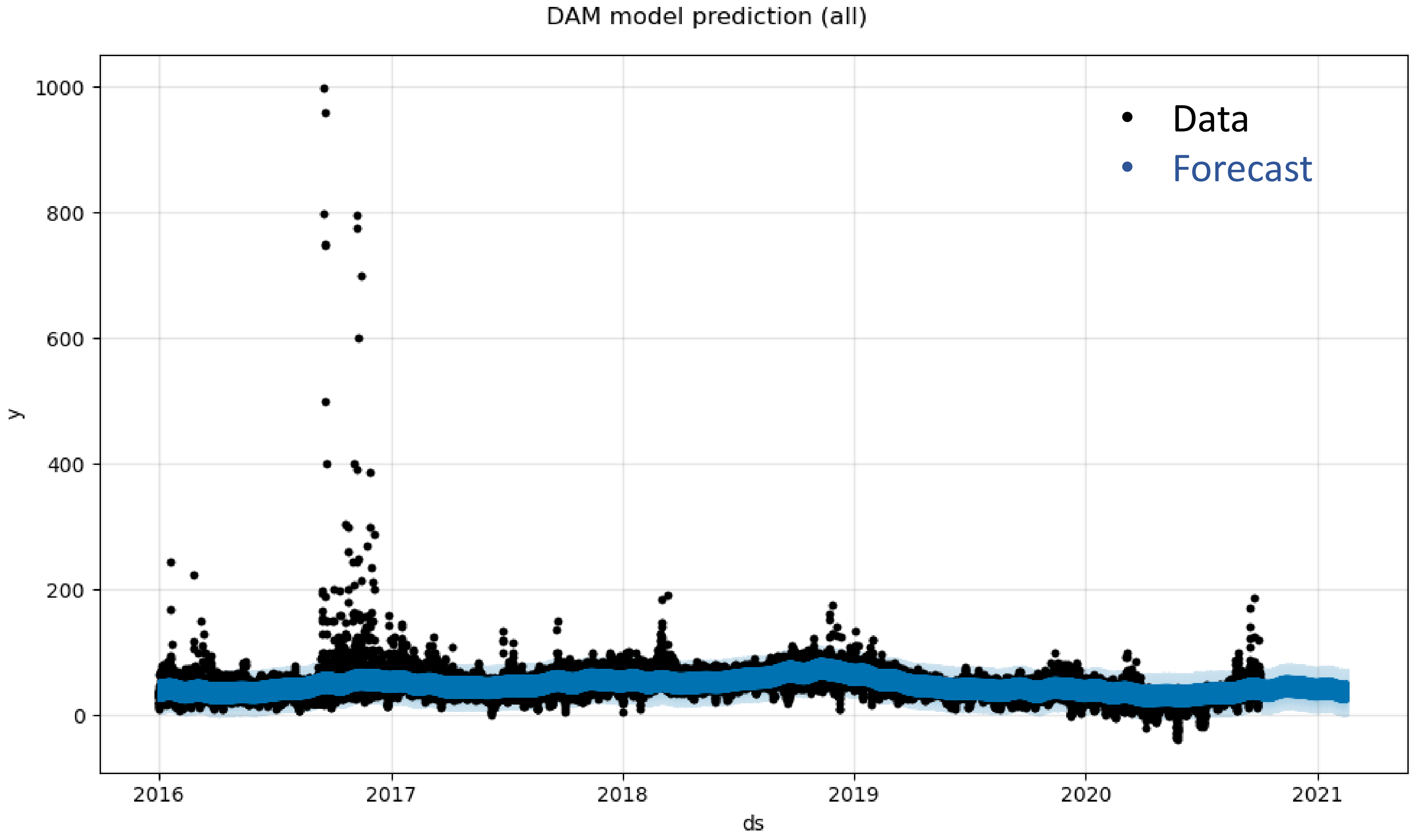
Train a time series model, using fb_prophet

- Split off the last 3 months of data, as a test period, and train a model on the remainder.
- We are only training a model on the DAM series, due to time constraints.

- Using the default settings, we find three components of interest.
  - Trend component: The underlying price movement, after stripping away the periodic components.
    - The movement in this component reflects the changes in the overall energy market over time, and appears to be quite large.
  - Yearly component: The repeated changes that happen yearly.
    - The amplitude is fairly low, only +/- 6 *units*
  - Daily component: Reflects the daily cycle pointed out on the previous slide.
    - The amplitude is larger than that of the yearly component.
- The sum of these three components is the predicted DAM price.



Predictions using the prophet model



- The prophet model provides a forecast that can be compared to the true data.
- The model does not capture all the variation in the true prices, but fits the overall trend

Final Thoughts / Next Steps

- The model is not great, when looking at the test period. However, the test period is the end of the dataset, and it is the spikiest period. The model likely does better for the smoother periods.
- MAE (mean absolute error) is probably the best metric for this data. I usually prefer MAPE, but it doesn't do well with true values near zero, which we have here. The negative r2 score is concerning, in the test period. It implies that this period is totally unlike the period used to train the model.

Ideas for Next Steps

- Clean the data a bit. De-spike the high spikes, perhaps. I would need to consult a domain expert to determine if this is a good idea. I don't like to remove outliers without understanding why they are there.
- Tune prophet model parameters. For example, we can set the initial strength of the seasonal components. Also, prophet has the capacity to fit holidays and additional regressors, which we are not using here.
- Do the same with the SIP data.
- Stretch Goal: Create an optimal control model to buy and sell energy throughout the day.
    - If you have capacity to buy energy, store it, and sell it later, this could be quite profitable. ::wink::
    - Use the predicted prices to find good control inputs, and then see how those inputs would perform on the actual data. Iterate as needed.

```
DAM Training Metrics
mean_absolute_error: 6.910045990562776
mean_absolute_percentage_error: 25274251034332.344
mean_squared_error: 237.5107104845655
r2_score: 0.4090536190402515

DAM Test Metrics
mean_absolute_error: 19.746258387210897
mean_absolute_percentage_error: 42603304770586.586
mean_squared_error: 2771.2711598942105
r2_score: -0.04077551497373699
```