14.05.2025

Justus Purat & Alexander Kammeyer
Software Project Distributed Systems

Consumption Data Forecast for HPC Systems

Sprint 1
Preprocessing

M. Ch
A. Er
A. Huth
M. Karn
Y. Kaya
M. Zent

Institute for Computer Science, FU Berlin

Digital Twin

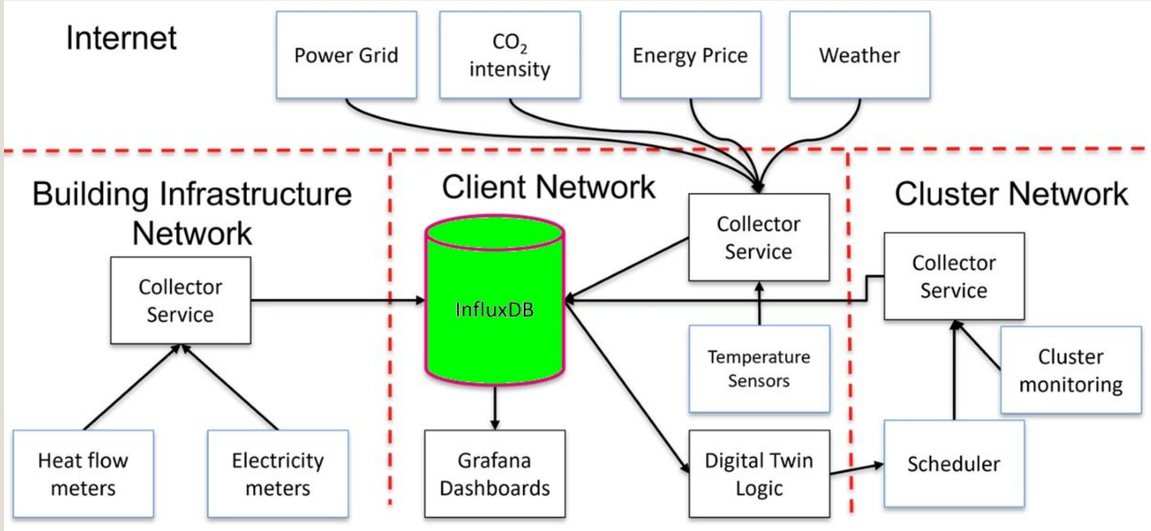
The Digital Twin is a virtual construct that represents a physical counterpart, integrates several data inputs with the aim of data handling and processing, and provides a bi-directional data linkage between the virtual world and the physical one.

- Physical counterpart : HPC cluster of PTB
- Data inputs : Cluster monitoring, Webservices
- Aim : Optimization of cluster usage, regarding e.g. more FLOPS per Watt, less energy per task, overall energy costs, CO2 emissions

Kammeyer et al.:

- 2023, Optimization of Energy Efficiency of an HPC Cluster, SMSI 2023, pp. 378-379
- 2023, Towards an HPC cluster digital twin and scheduling framework for improved energy efficiency, ACSIS 35:265-268
- 2024, HPC operation with time-dependent cluster-wide power capping, ACSIS 39:385-393
- 2024, Developing a digital twin to measure and optimise HPC efficiency, Measurement: Sensors, 2024.101481

Data Collection



[1]

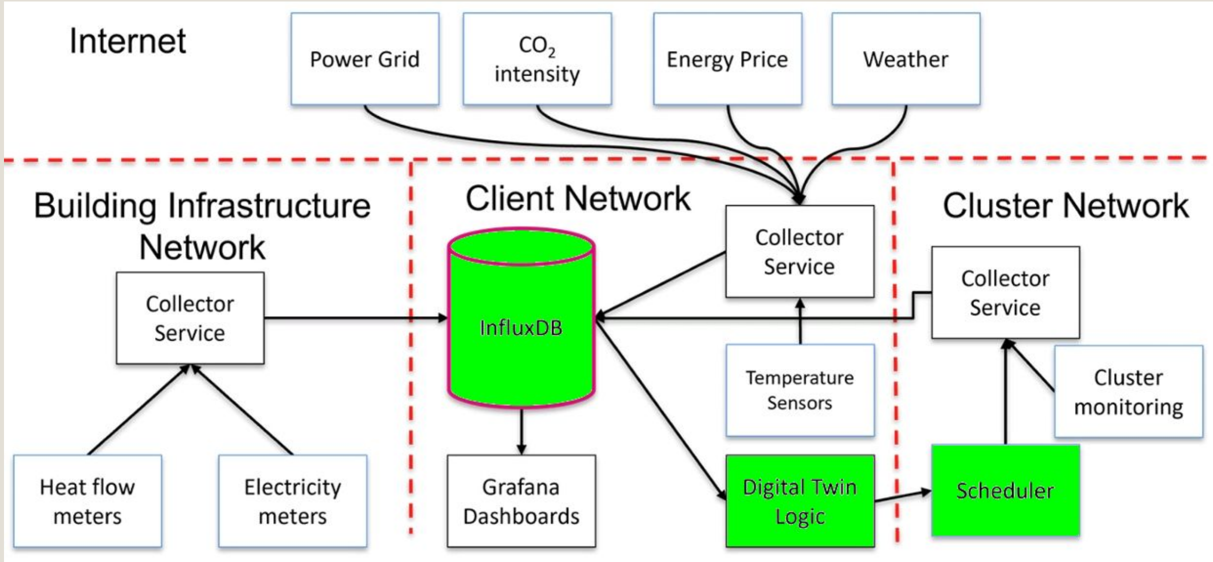
Data stored with InfluxDB, a time series database system

- Accessible via web API
- Queries can be formulated in SQL-alike InfluxQL or the functional script language Flux

Forecasting

Problem : Limited budget of money, energy, and CO_2 emission

Task : Forecast energy prices and CO_2 footprint for effective scheduling



What has been done

- Created the GitLab for our project
- Created the database connection on our local system
- Queried the database to get familiar with structure and shape
- Analyzing the data with python pandas library
- Creating Charts
- Analyzed results of SWP 2023/24 group

Database Scheme

General Bucket Columns

Column	Meaning
_start, _stop	Provided time interval within the requested one
_time	Field timestamp within the provided interval
_field	Field key, designating a category of measurement
_value	Field value
table	Field index. Can change with each query

Database Scheme

Bucket: co2

Column	Possible values
_measurement	co2
_field	carbonIntensity, fossilFuelPercentage
contryCode	DE, PL, FR, …
status	ok

Database Scheme

Bucket: co2

Column	Possible values
_measurement	co2_calculated (based on bucket “energy”)
_field	carbonIntensity, fossilFuelPercentage
contryCode	DE

Database Scheme

Bucket: energy

Column	Possible values
_measurement	energy_production
_field	“Kernenergie”, “Biomasse”, “Wind Onshore”, “Photovoltaik”, “Braunkohle”, “Steinkohle”, “Erdgas”, “Pumpspeicher”, “Wasserkraft”, “Sonstige Erneuerbare”, “Sonstige Konventionelle”, “Wind Offshore”
region	DE

Database Scheme

Bucket: price

Column	Possible values
_measurement	price
_field	“Deutchland/Luxembourg” (Sic! Typo from the web service)
region	DE

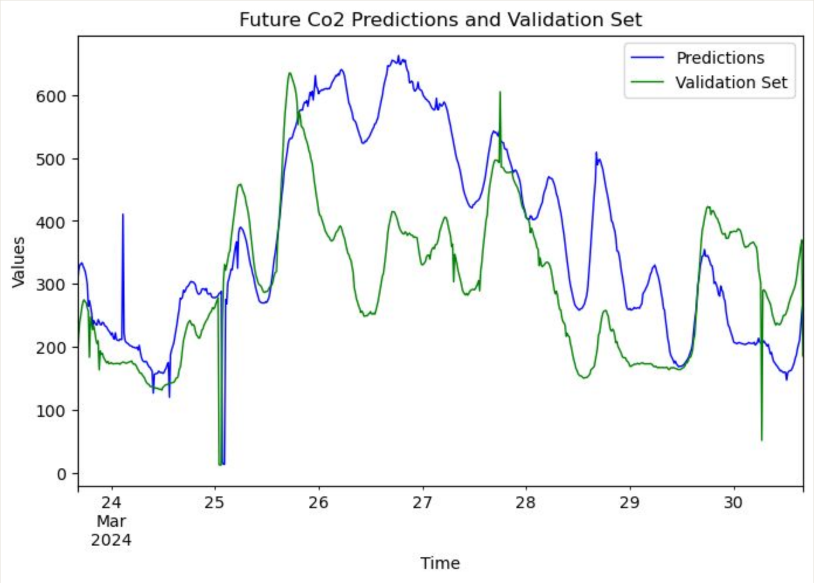
Database Scheme

Bucket: weather

Column	Possible values
_measurement	weather
_field	icon, pressure_msl, solar, source_id, sunshine, temperature, visibility, wind_direction, wind_gust_speed, wind_speed

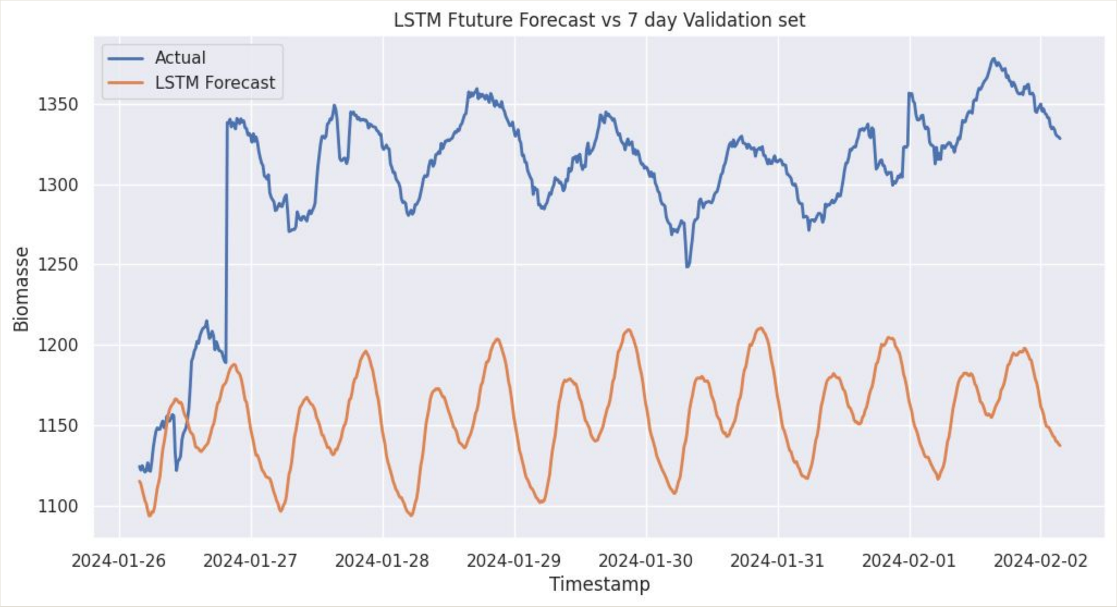
SWP WS 2023/24
Results

Forecast on the Carbon Intensity, calculated from the energy mix data, using a Random Forest Regressor trained on the Exponential Moving Average, with a Mean Absolute Error of 51%.



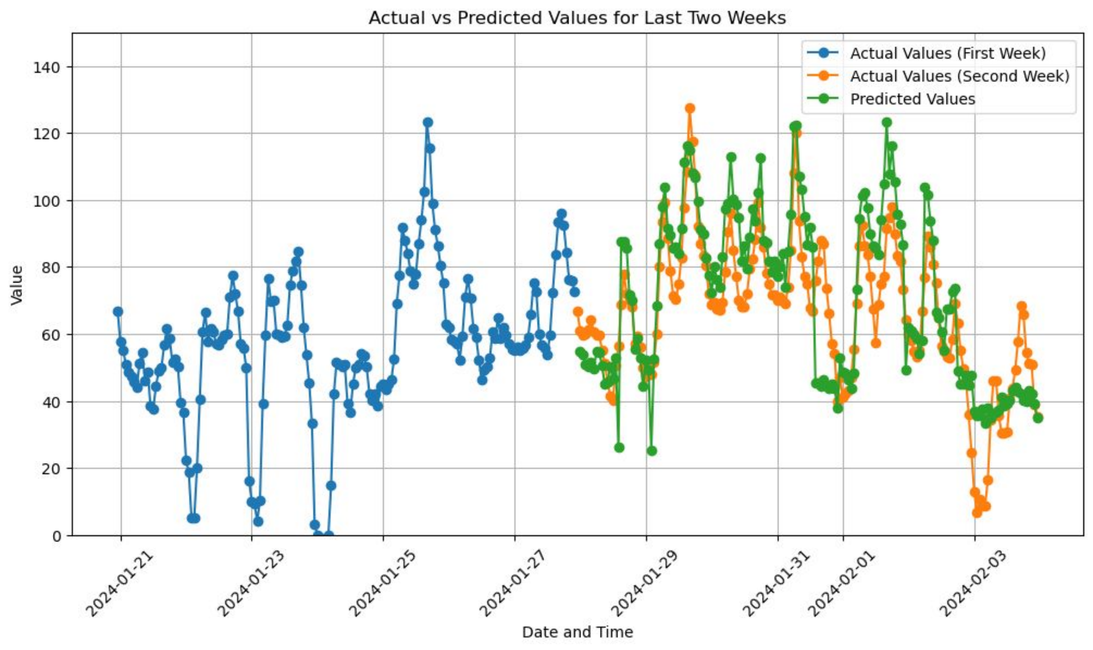
SWP WS 2023/24
Results

Forecast on the production of each energy carrier, using a Long Short-Term Memory Neural Network. Mean Absolute Error: 9-505%



SWP WS 2023/24
Results

Forecast on the energy price with two Random Forest Approaches.
Mean Absolute Error of best approach: 25%



SWP WS 2023/24

Issues

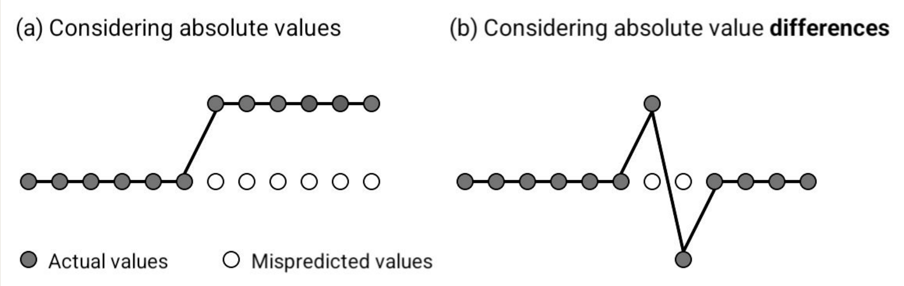
- Random Forest Tree produces more or less correct trends, but oftenly exaggerates
- Long Short-Term Memory NN is too static and reacts bad on any short- and mid-term deviation from long-term trends
- In general bad reaction on unexpected events, e.g. the shut-down of nuclear energy

- No use of correlations made
- Unreasonable to forecast each energy carrier for itself instead of grouping into fossiles and renewables

SWP WS 2023/24
Lessons for 2025

- Find and exploit meaningful **Correlations**, i.e. with causality

- **Predict Trends**
  - Abolute values can become easily obsolete, e.g. by political choices, but relative changes are usually more stable
  - Ruptures in absolute data with long-term impact are only outliers, when looking at differences



Data Visualization

Table for CO2 intensity and Fossil Fuel Percentage

CO_2-Daten (letzte 6 Monate)

CO_2-Intensität (gCO_2/kWh)	Fossiler Anteil (%)	Zeit
652.7663525109017	72.23472593426173	2024-11-12T04:15:00+00:00
652.6632070214545	72.78257639082382	2024-11-12T04:30:00+00:00
653.1047392932152	73.4248013880011	2024-11-12T04:45:00+00:00
652.3043517191977	73.36362822349571	2024-11-12T05:00:00+00:00
650.7329293461555	73.47516796206034	2024-11-12T05:15:00+00:00
646.0562019241071	73.23605665188232	2024-11-12T05:30:00+00:00
640.5154828862358	72.3294373471052	2024-11-12T05:45:00+00:00
635.5781599837167	71.11337268471402	2024-11-12T06:00:00+00:00
633.6506722438004	70.54277462404143	2024-11-12T06:15:00+00:00
630.9713368313885	70.15036416320777	2024-11-12T06:30:00+00:00
628.2613522085558	70.05492380289317	2024-11-12T06:45:00+00:00
621.603066633733	69.34506279569075	2024-11-12T07:00:00+00:00
618.8232682394552	69.03525604846295	2024-11-12T07:15:00+00:00
616.886651247626	69.04157499858971	2024-11-12T07:30:00+00:00
615.5335762721415	69.03845070157972	2024-11-12T07:45:00+00:00
607.5458911419423	67.81547860008097	2024-11-12T08:00:00+00:00
606.5950408659156	68.06199838008983	2024-11-12T08:15:00+00:00
605.394631167831	68.175733905422	2024-11-12T08:30:00+00:00
603.5312465637943	68.19814536524575	2024-11-12T08:45:00+00:00
606.2502120441052	68.51606003613969	2024-11-12T09:00:00+00:00
606.1563402052257	69.16587518855802	2024-11-12T09:15:00+00:00
604.2923329233292	69.13041857691304	2024-11-12T09:30:00+00:00
602.6317211816233	68.99450191035318	2024-11-12T09:45:00+00:00
600.9328917263267	68.41095283784782	2024-11-12T10:00:00+00:00

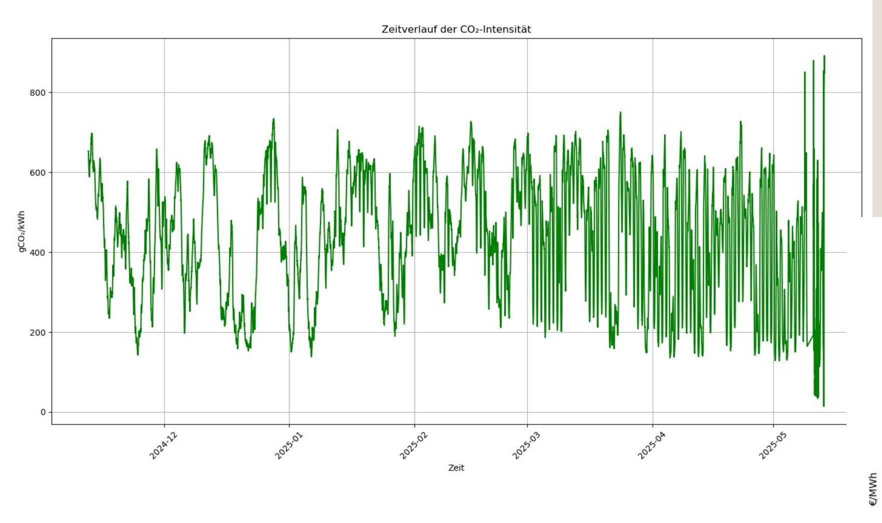
Data Visualization

Table for energy types and price

Energieproduktion & Energiepreis (letzte 6 Monate)

time	Biomasse	Braunkohle	Erdgas	Photovoltaik	Sonstige Erneuerbare	Sonstige Konventionelle	Steinkohle	Wasserkraft	Wind Offshore	Wind Onshore	tchland/Luxembo
.1-12T04:15:00-	1008.25	2891.25	2813.75	1.75	32.5	270.75	1727	333.75	409.25	1165	118.01
.1-12T04:30:00-	1010	2890.75	2949.75	1.75	32.5	271	1725	337	387.25	1159.5	118.01
.1-12T04:45:00-	1018.75	2933.5	3128.5	2	32.5	270.75	1708	334	371.25	1151.25	118.01
.1-12T05:00:00-	1043.5	2989.75	3216.25	1.5	32.5	284.75	1702.5	343.25	348.25	1171.75	147.67
.1-12T05:15:00-	1058.25	3016.75	3353.75	1.5	32.25	280.5	1715.25	367	331	1178.75	147.67
.1-12T05:30:00-	1071.25	3018.5	3517.25	1.5	32	269.5	1739.75	411.75	312.5	1202.25	147.67
.1-12T05:45:00-	1088.75	3019.25	3610.5	1.75	32	269.75	1748.75	448.75	304	1216	147.67
.1-12T06:00:00-	1116.75	3047.25	3689.5	2	32	275	1722.75	474	286.75	1210.5	185.52
.1-12T06:15:00-	1131	3088.75	3750.5	2.5	32	289	1725.75	484.75	279.25	1226	185.52
.1-12T06:30:00-	1135.75	3101.25	3828.5	12	32	304.25	1723.5	494	268.25	1242	185.52
.1-12T06:45:00-	1142	3095	3916.5	49.25	32	325	1719.5	486	265.75	1241.5	185.52
.1-12T07:00:00-	1153.75	3098.25	3994.75	113	32	340.75	1690.5	520	250.5	1249.75	201.64
.1-12T07:15:00-	1151.25	3102	4038	205.25	32	339	1694.75	499.25	238.25	1249.75	201.64
.1-12T07:30:00-	1152	3103.5	4040.75	318.25	32	332	1703	495.5	225.25	1226.75	201.64
.1-12T07:45:00-	1150.5	3104.25	4051.25	443.5	32	327.75	1705.25	460.25	227	1183.25	201.64
.1-12T08:00:00-	1139	3104.75	4092.5	573.5	32	333.5	1683	524.75	212.5	1143.25	223.33
.1-12T08:15:00-	1139.5	3103.25	4102	700	31.75	354	1684.25	513.5	200	1092.75	223.33
.1-12T08:30:00-	1138	3104	4127.5	830.25	32	374	1689.75	492.5	190.25	1045.5	223.33
.1-12T08:45:00-	1135	3106	4129.25	947.5	32	374.75	1693.25	480	187	1017.5	223.33
.1-12T09:00:00-	1098.75	3110.75	4099.25	1058.75	32	373.5	1706.25	438	156.75	972.75	190.24
.1-12T09:15:00-	1091.75	3103	4096.75	1164	32	375.25	1710	444.5	128.25	945.5	190.24
.1-12T09:30:00-	1083.25	3112.75	4072	1260	31.75	375	1713.75	465.75	116.5	928	190.24
.1-12T09:45:00-	1078.75	3113.25	4047.75	1360	32	377	1716.75	445.25	113.75	912.75	190.24
.1-12T10:00:00-	1057.25	3117.5	4003	1435.75	32.25	378.5	1770	450.75	115	928.25	171.08

Illustration of the relations



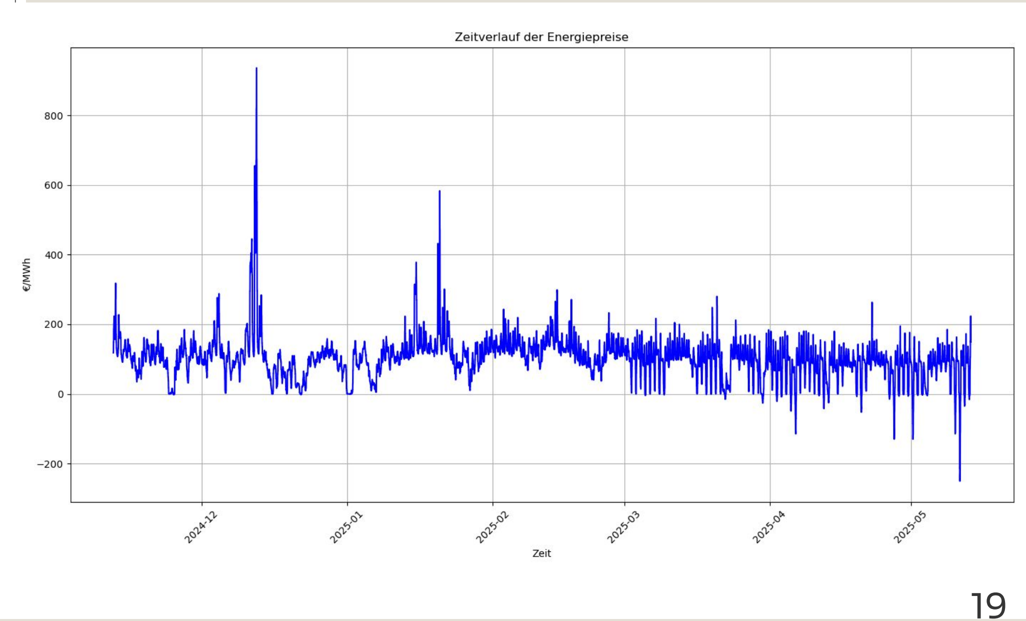


Illustration of the relations

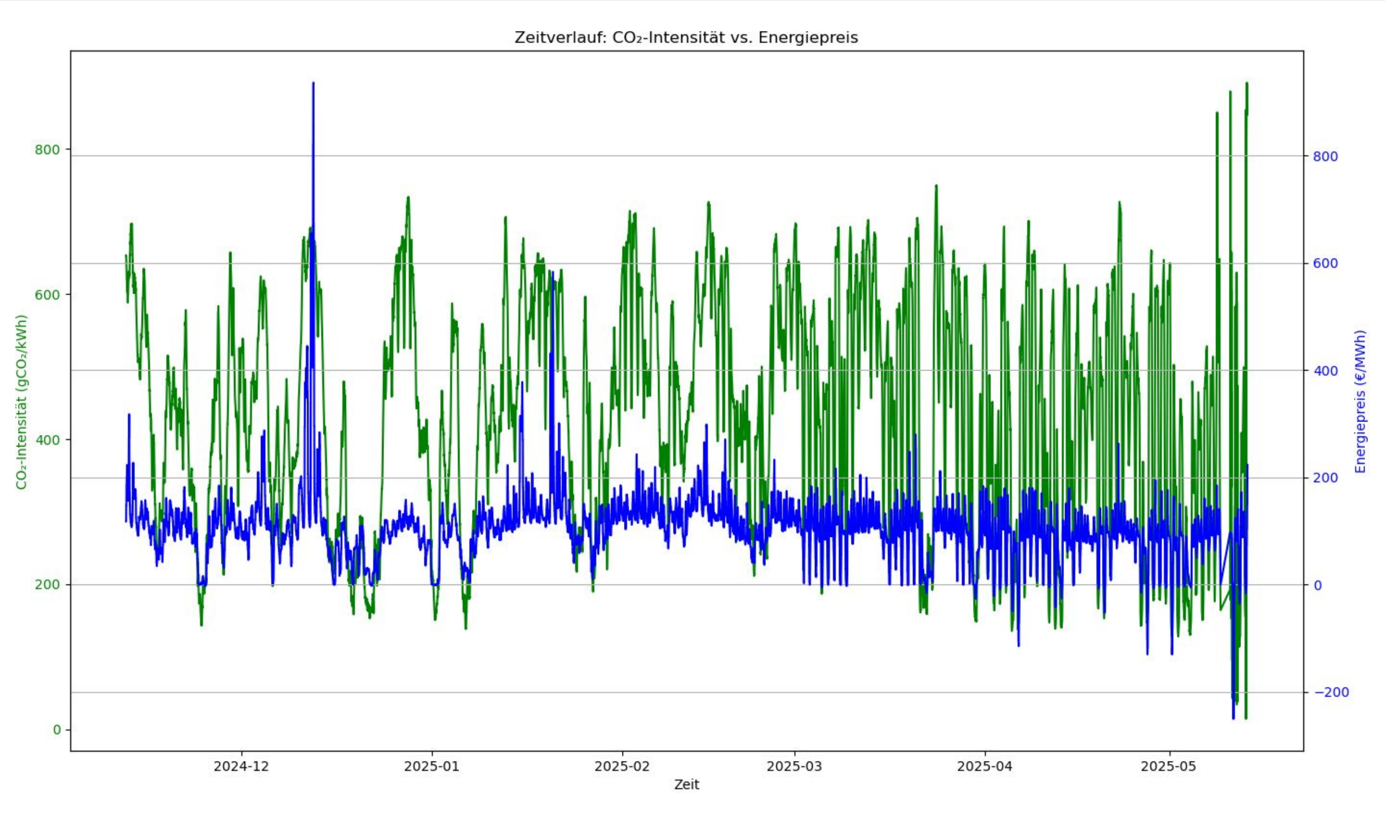


Illustration of the relations

Fossile energy

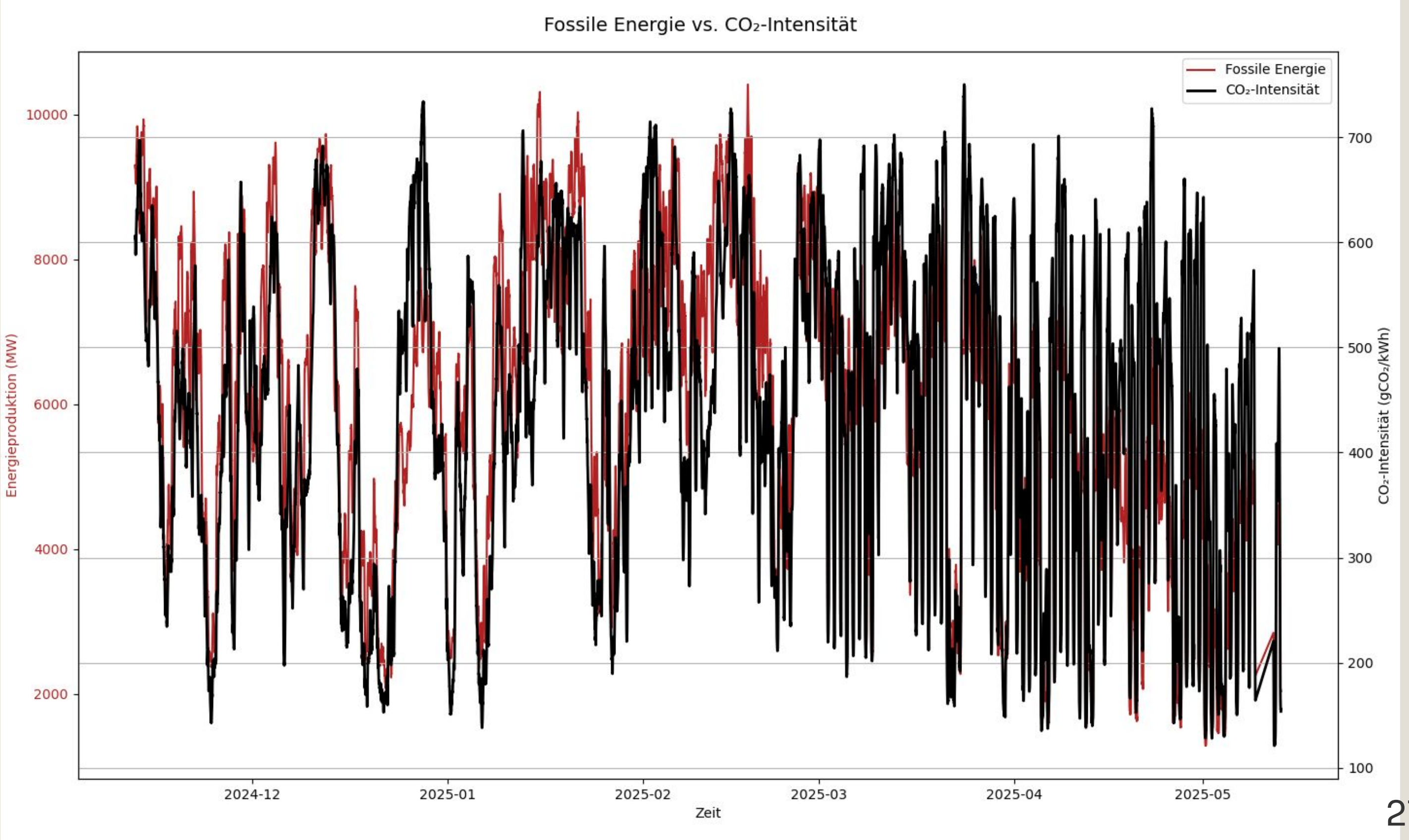
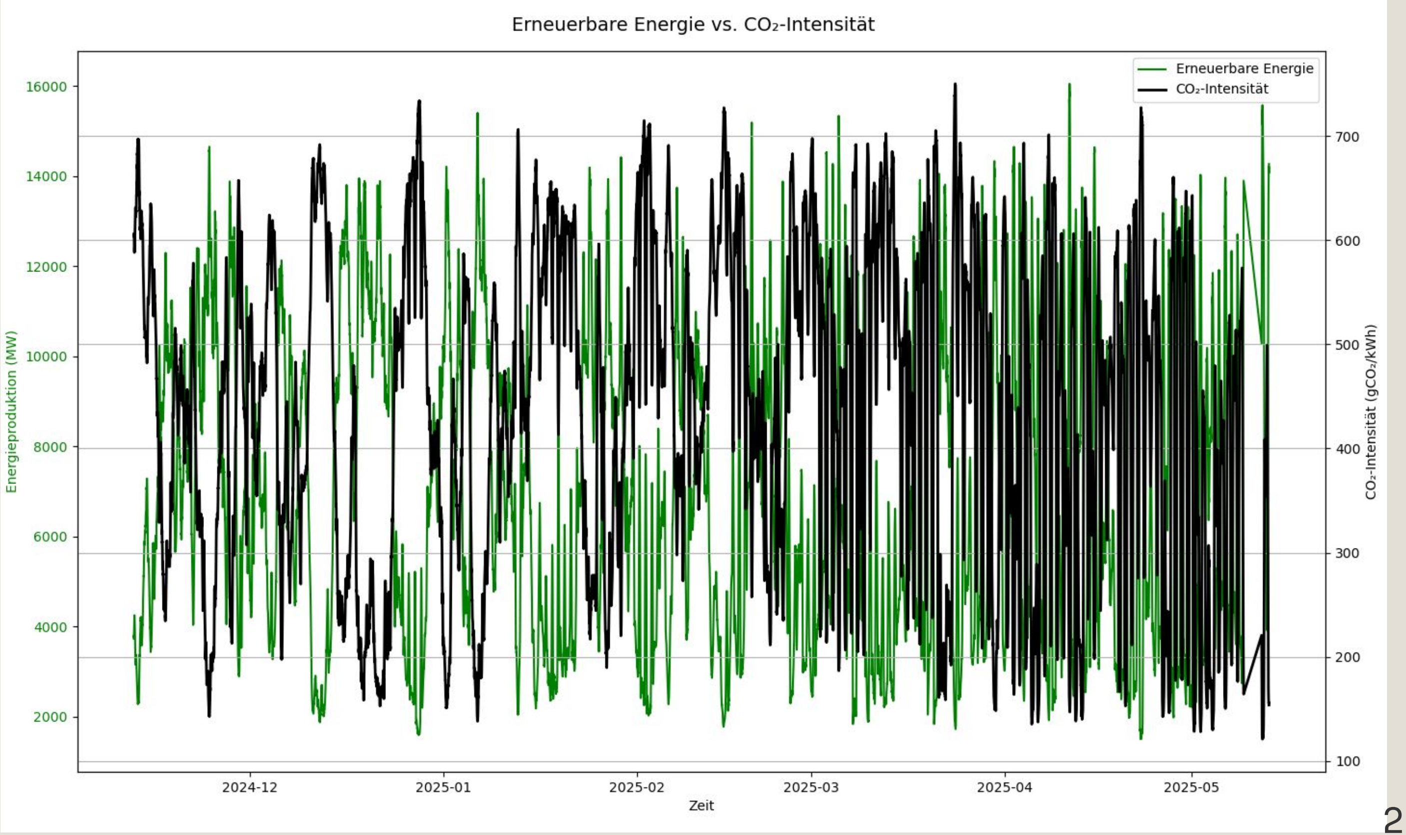


Illustration of the relations

Renewable energy



Did unexpected problems occur?

- Communication in the group
- Handling of gaps in the data

Previous Timeline

	Energy Price Group	CO2 Group	Dates
Sprint 1	Data prep & EDA; benchmark ARIMA	Data prep & EDA; energy-mix analysis	30 April – 14 May
Sprint 2	Train ML models (TimesFM, XGB); feature engineering	Train ML models; incorporate external regressors	14 May – 28 May
Sprint 3	Develop LSTM; hyperparameter tuning	Time-series cross-validation; tune ML and simple RNN models	28 May – 11 June

Outlook

- Meaningful correlations

- Data handling
  - Dealing with data gaps
  - Cleaning outliers

- Making decisions about models
  - Continue with Random Forests? Exclude LSTMs?
  - Other models?

Thank You

References

[1] A. Kammeyer et al. (2024), *Developing a digital twin to measure and optimise HPC efficiency, Measurement: Sensors*, 2024. 101481