The patterns and forecasting of household instantaneous demand using a one-year sample dataset

Dr. Xuezhi Zeng

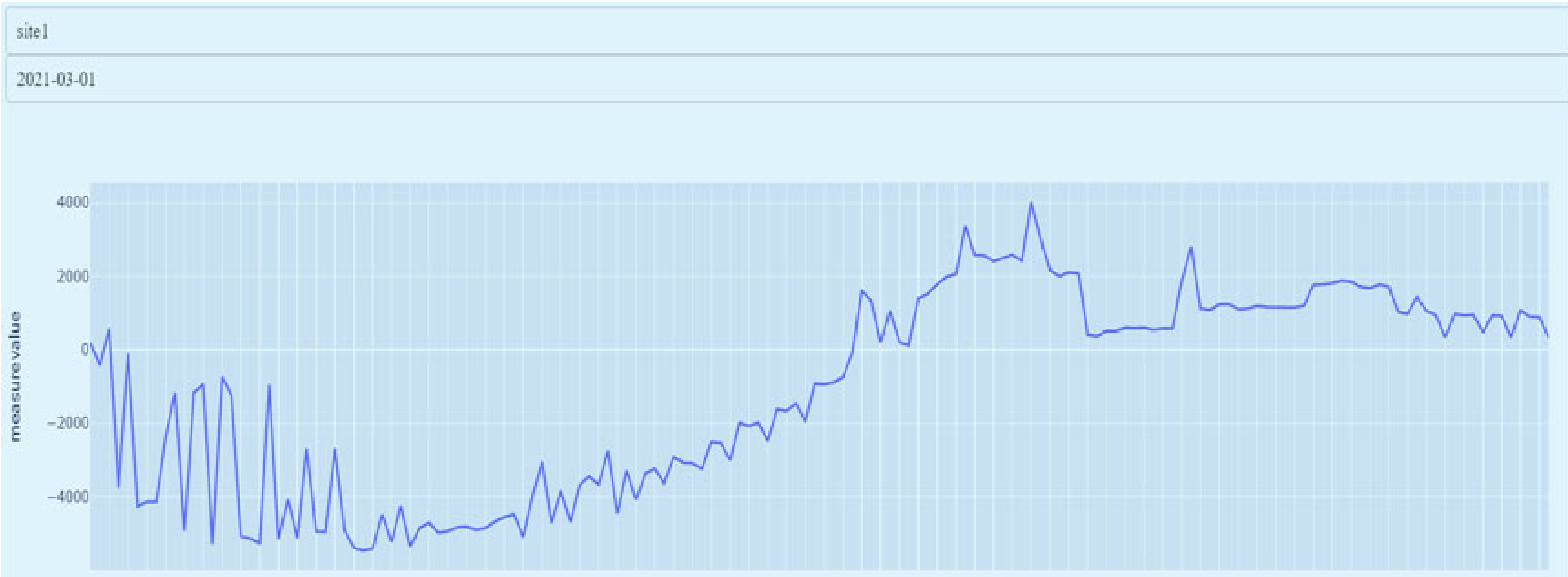
Content

- Dataset introduction
- Exploratory Data Analysis
- Forecasting of household instantaneous demand: two example cases
- What customer insights are provided by the analysis?
- Future improvement

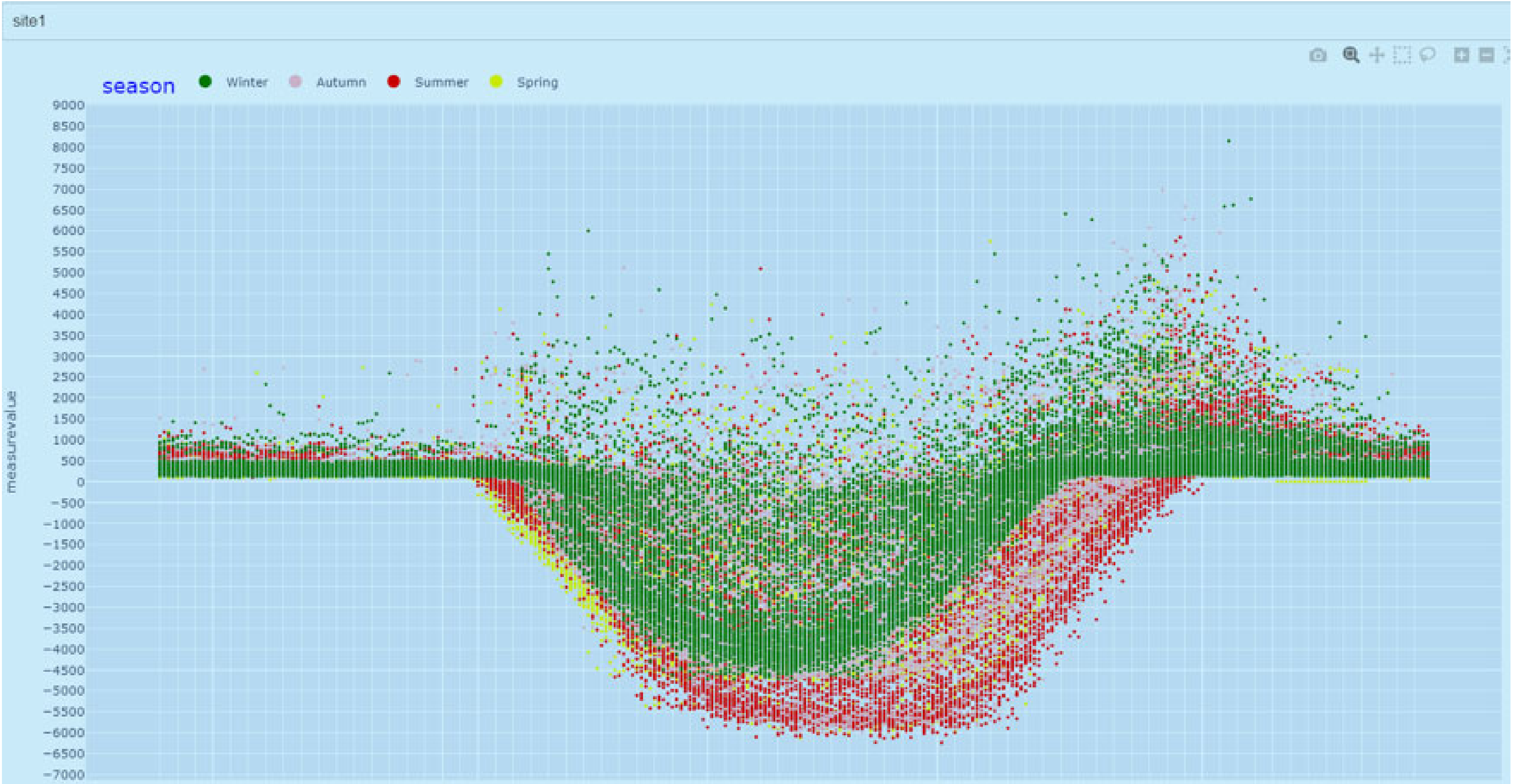
About the data

- 100 customers
- Time zone: XX
- Period: XXXX-XX-XX ~ XXXX-XX-XX
- Format: Apache parquet
- Rows: 10 millions +
- Time series interval: 5 minutes

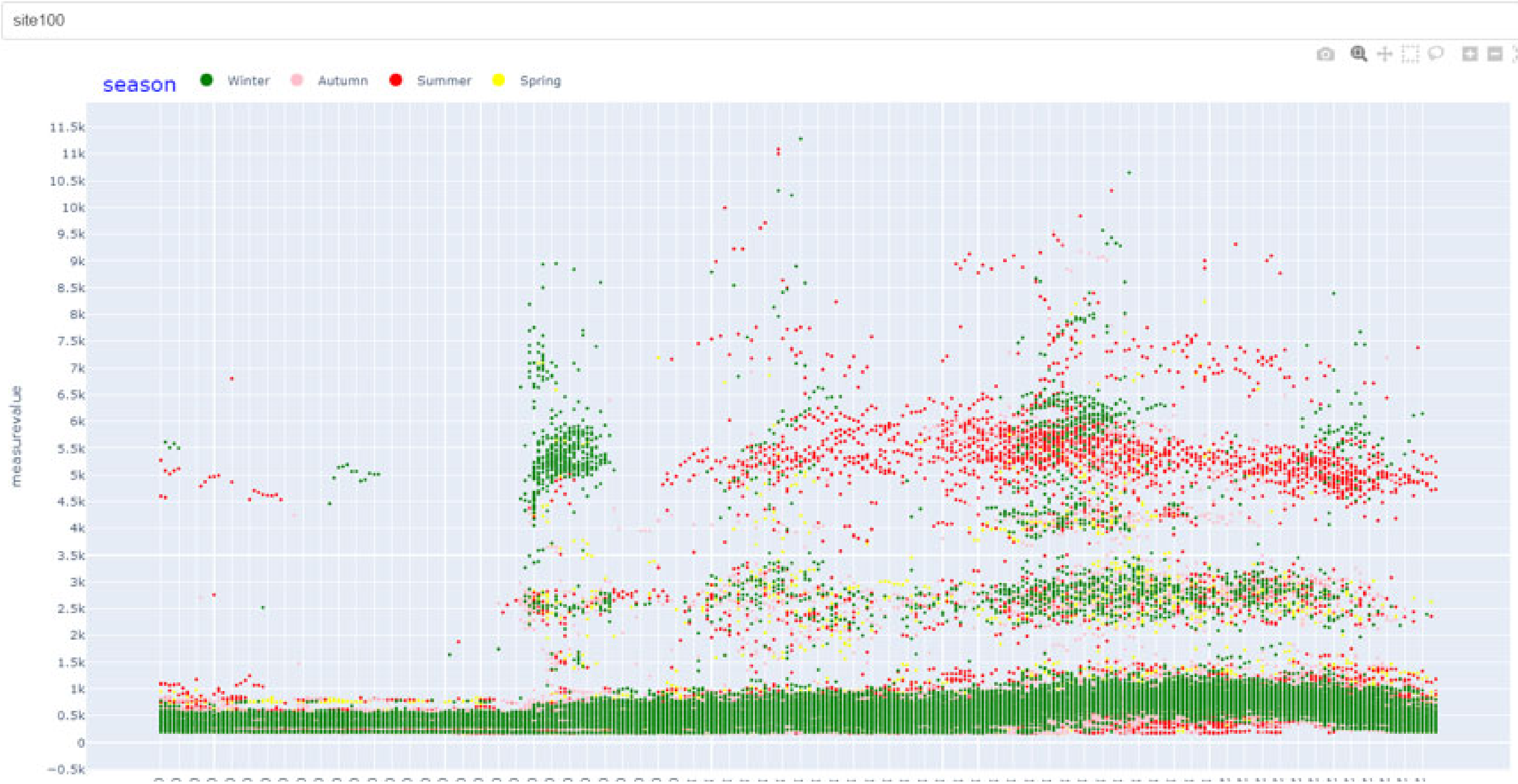
Understanding the patterns of instantaneous demand in a day for each household



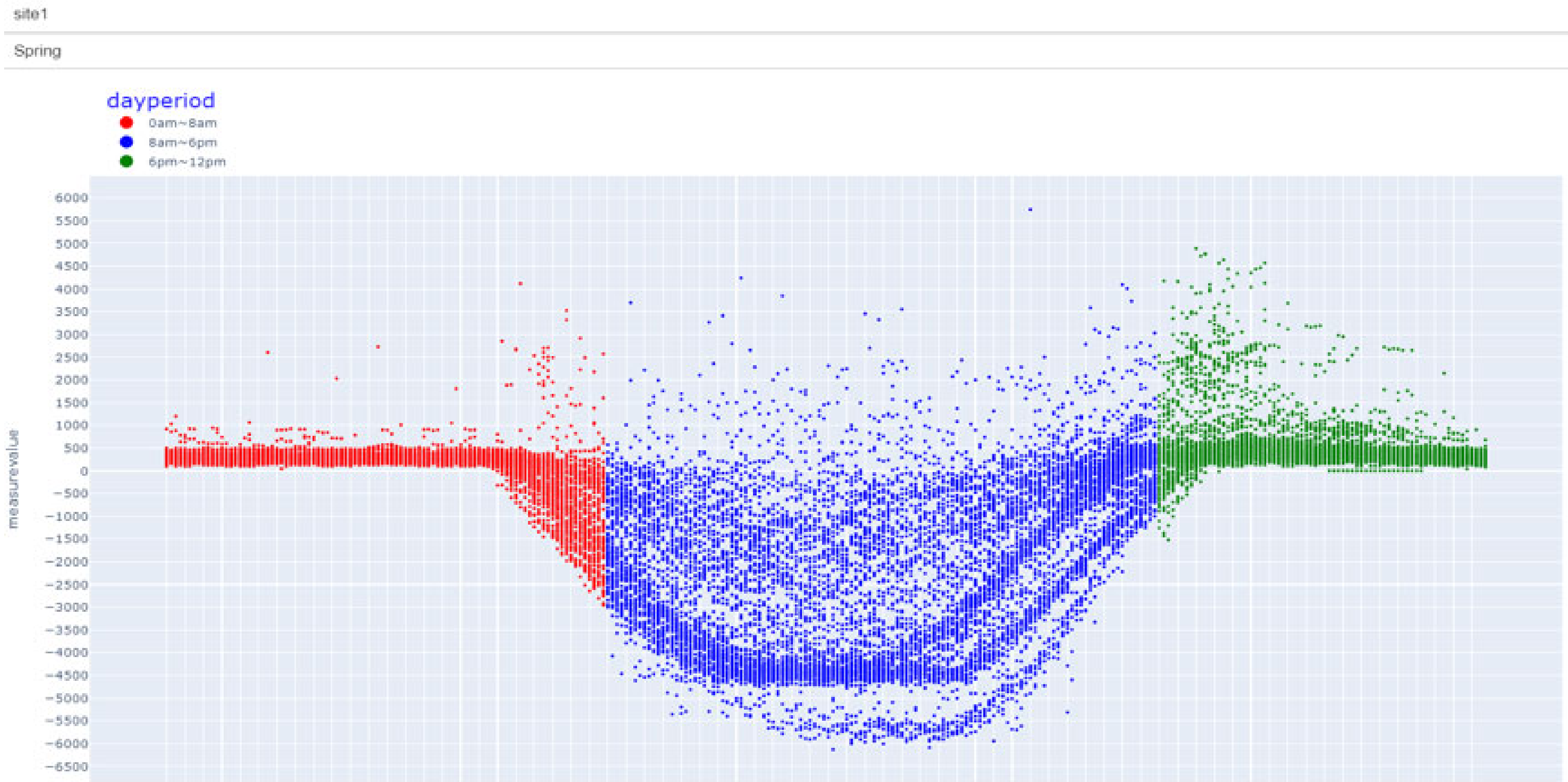
Instantaneous demand across seasons (solar-connected household)



Instantaneous demand across seasons (grid-dependent household)

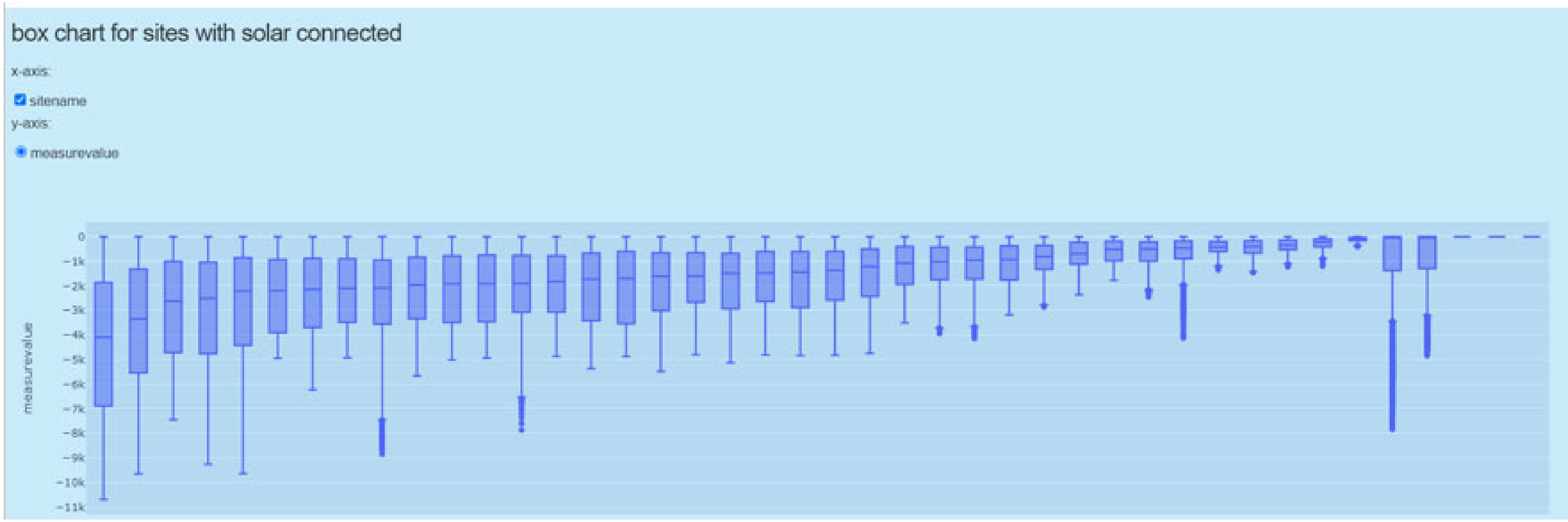


Instantaneous demand during day period (solar-connected household)

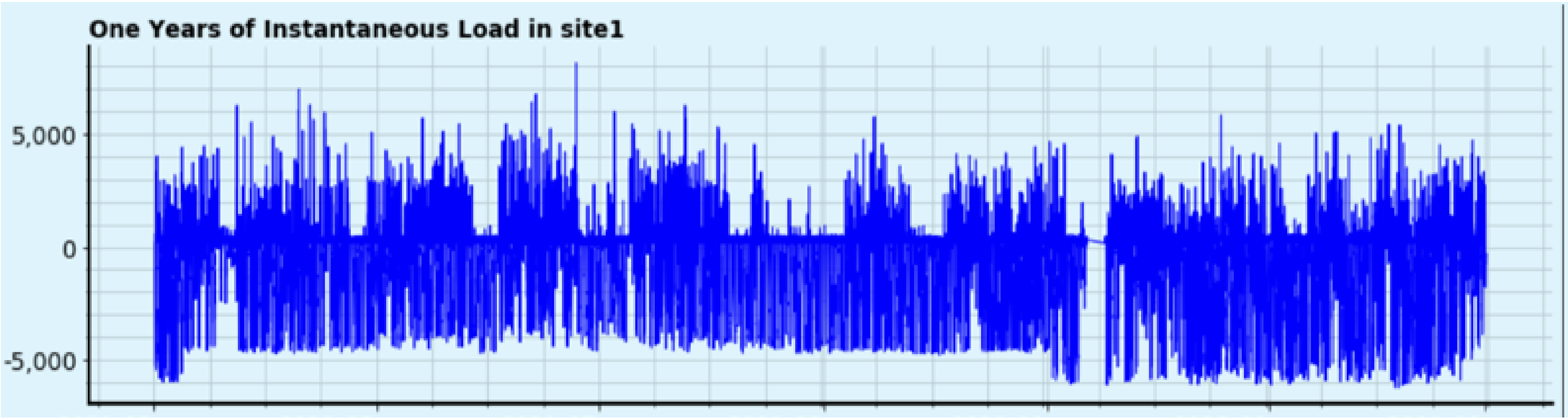


Instantaneous demand during day period (grid-dependent household)

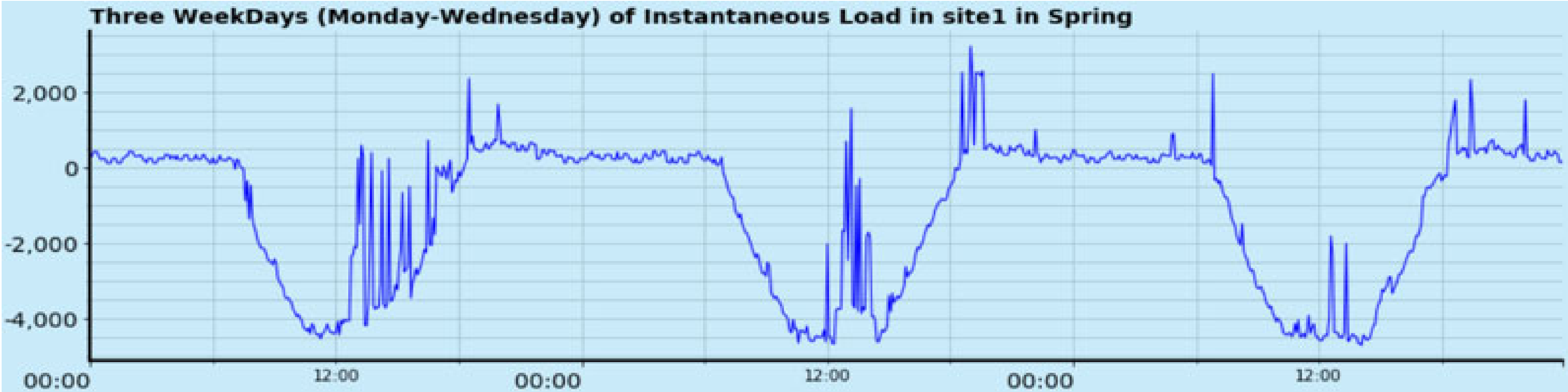
Solar insolation of solar-connected households – ranked by the median of solar insolation demand value

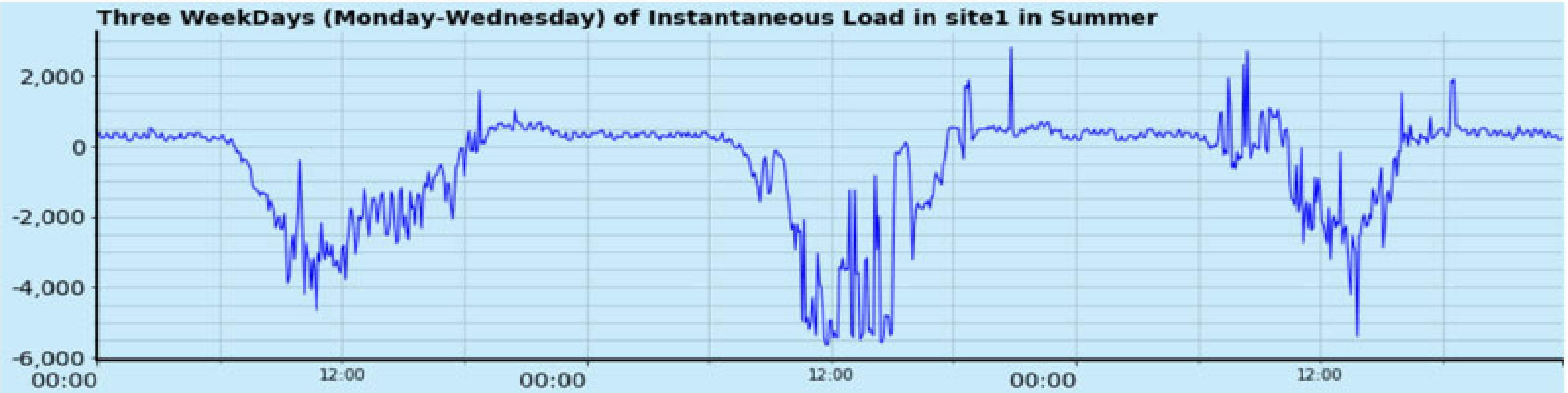


Example case I – Site 1: Instantaneous load over a year

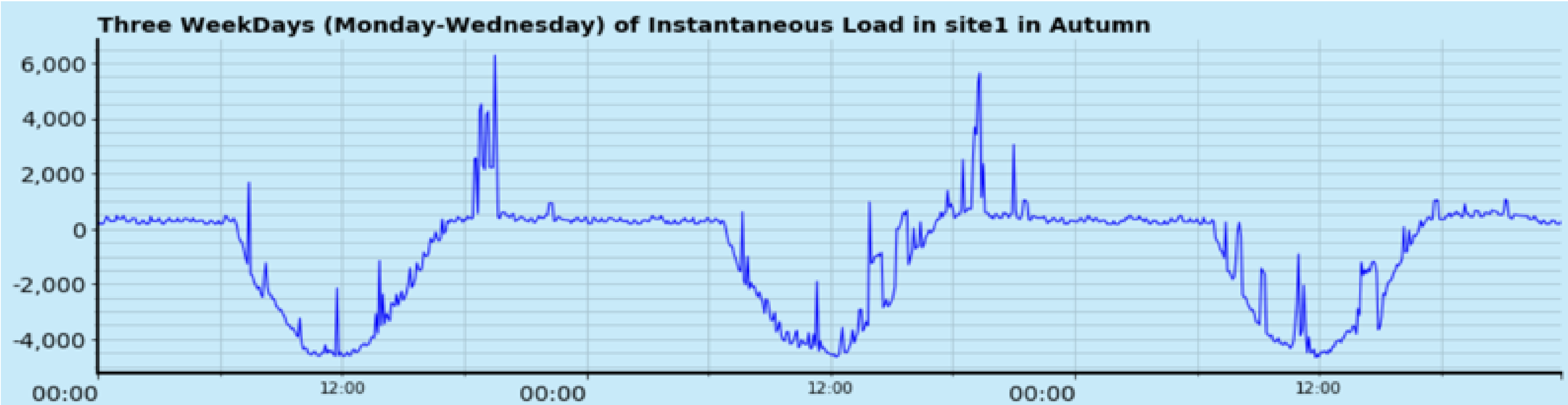


Example case I – Site 1: Instantaneous load of three weekdays by season

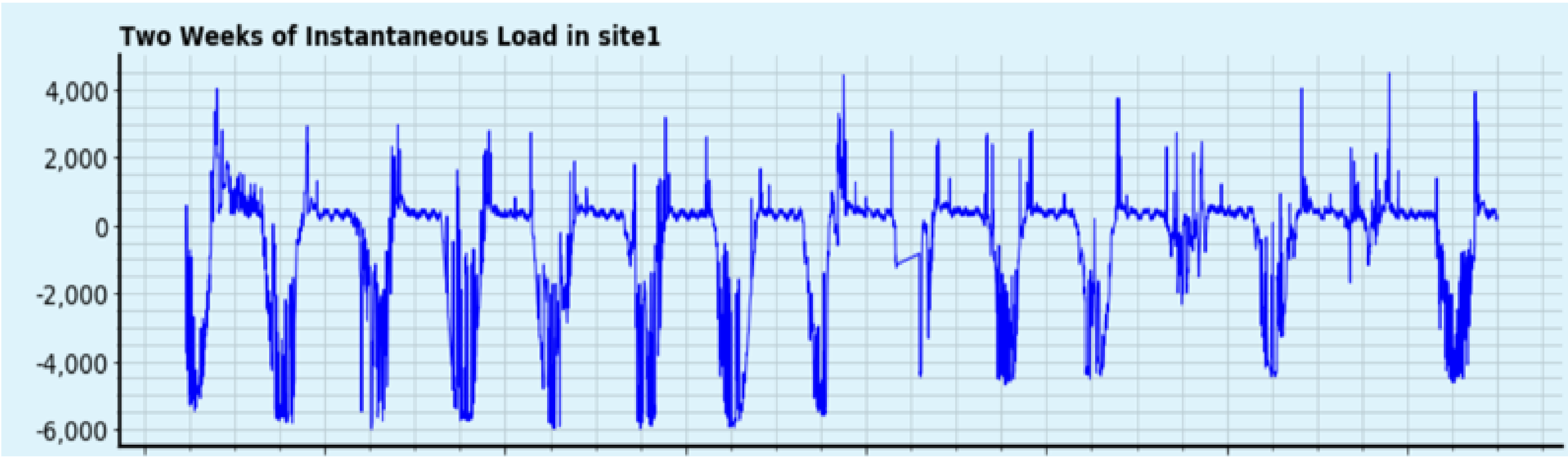




Example case I – Site 1: Instantaneous load of three weekdays by season



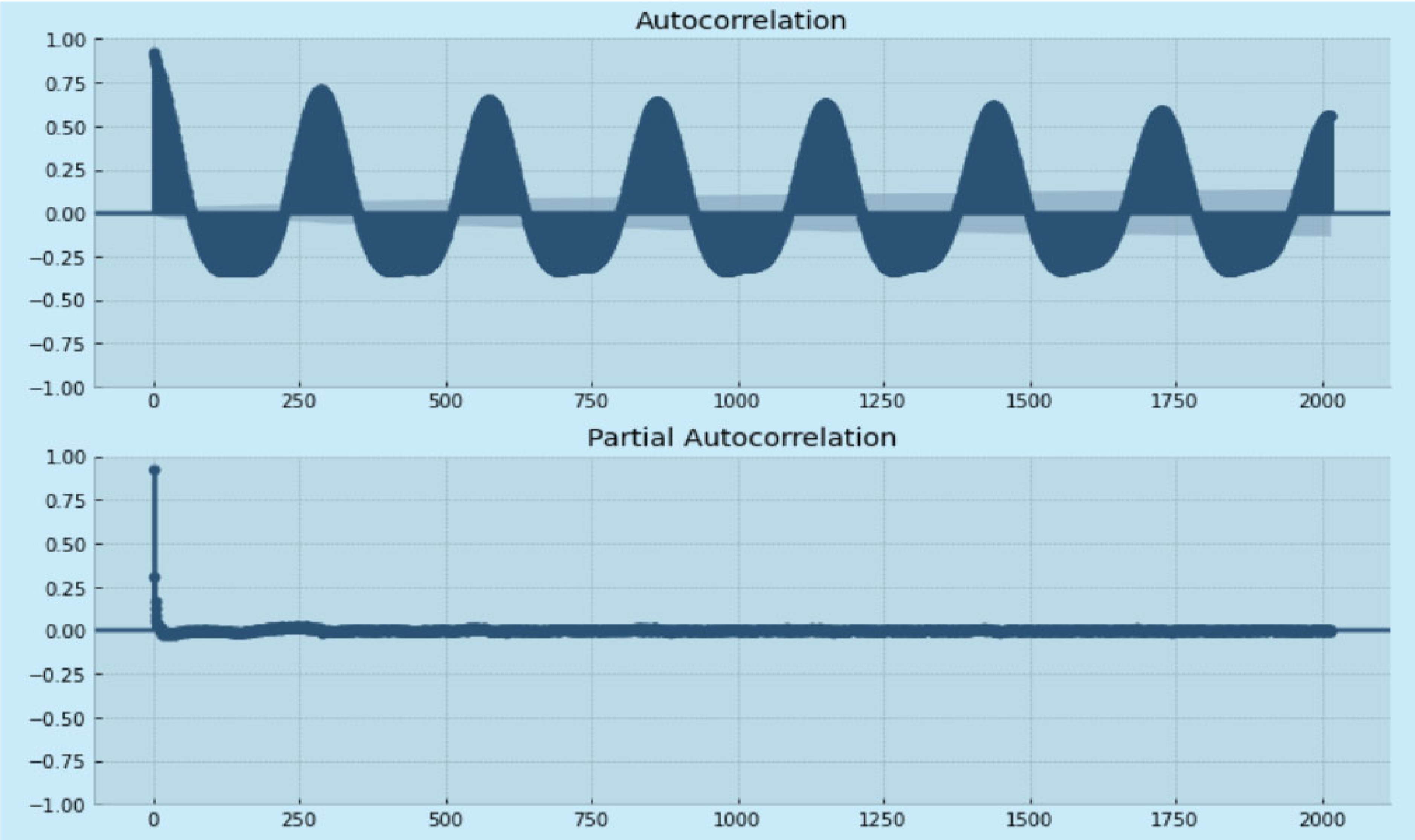
Example case I – Site 1: Instantaneous load of two weeks



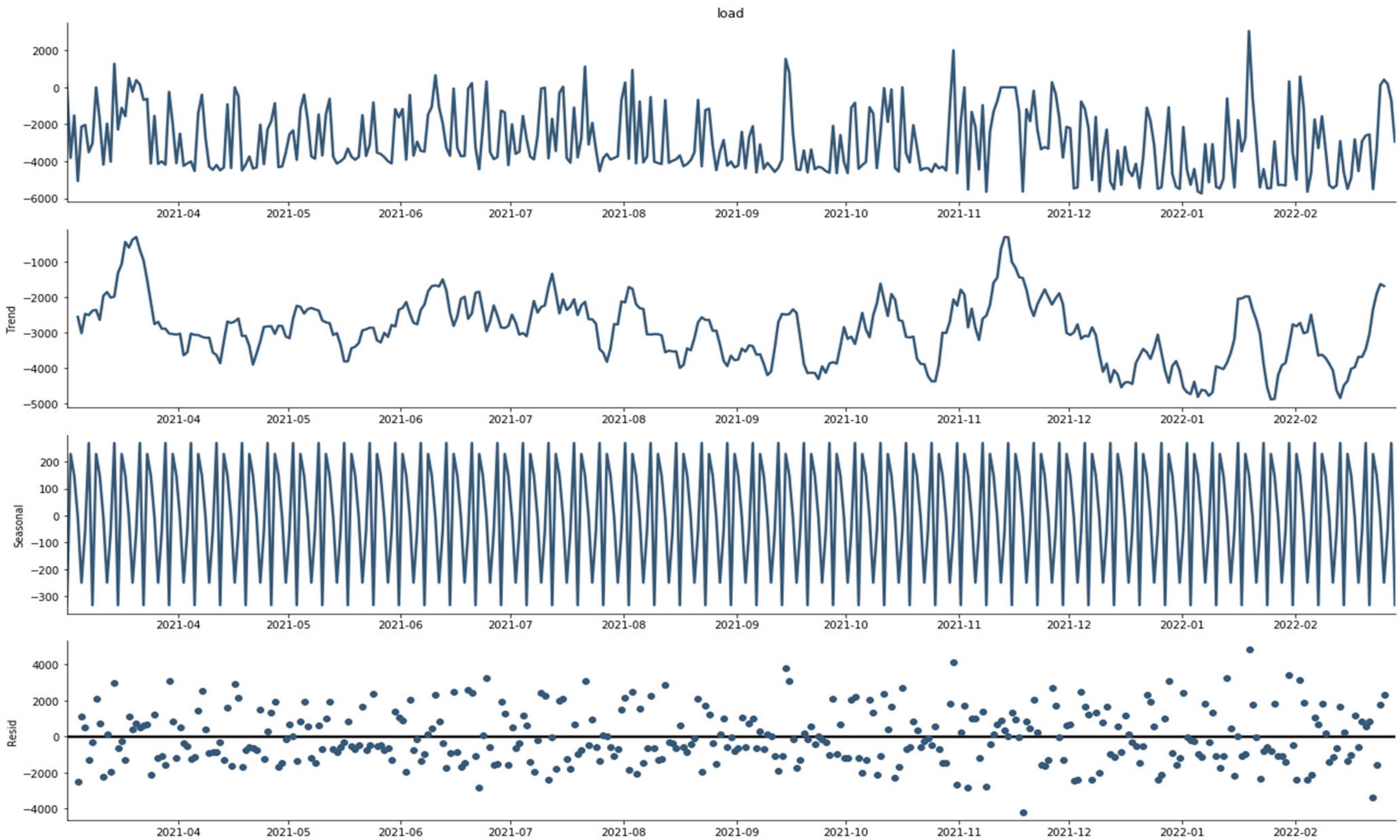
Example case I – Site 1: stationarity test

	Value	module
name	load-site1	nan
ADF_Statistic	-37.565288	adfuller
p-value	0.000000	adfuller
num_lags_used	67	adfuller
n_observations_used	103240	adfuller
IC_for_best	1619622.194613	adfuller
1%	-3.430413	adfuller
5%	-2.861568	adfuller
10%	-2.566785	adfuller
likely stationary	True	p-value < 0.05

Example case I – Site 1: autocorrelation test



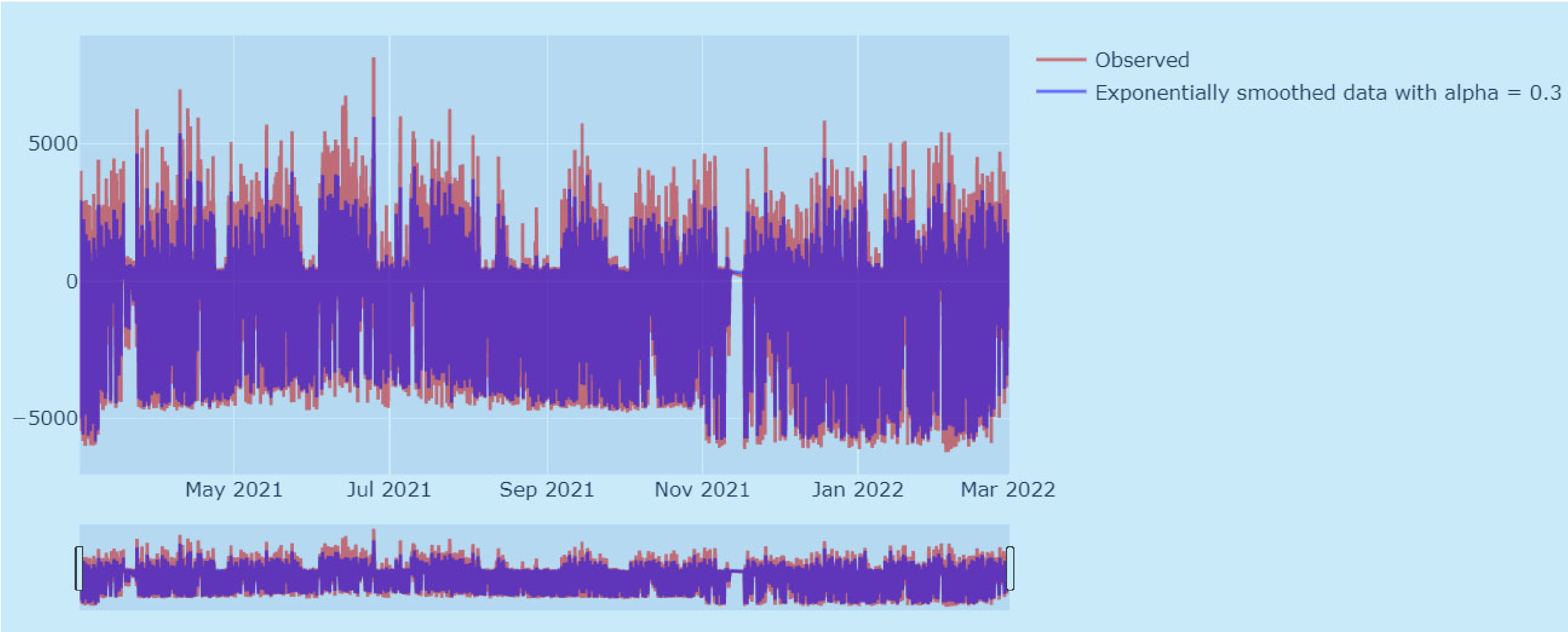
Example case I – Site 1: Decomposing trend, seasonality and noise



Performance comparison of forecasting models (Site1)

Forecasting Models	**Metrics**	
	RMSE	MAE
Baseline Persistent Forecast	1466.38	890.35
Linear Regression	1051.13 (train set) RMSE is very big in test set, indicating overfitting	701.58 on train set
Ridge Regression	1082.92 (train set) 1389.68 (test set)	737.41 (trainset) 1115.09 (test set)
Random Forest Regression (reduced feature dimension)	1184.73 (train set) 1497.81 (test set)	754.77 (train set) 999.43 (test set)
Random Forest Regression (reduced feature dimension, and adding lags as feature)	858.38 (train set) 1086.26 (test set)	504.56 (train set) 662.53 (test set)
Exponential smoothing	452.32	245.44

EWMA for site1



What customer insights are provided by the analysis (1)

- **Two groups of customers can be observed and classified according to the presence of Distributed Energy Resources**
- **Daily pattern**:
  - For solar-connected customers, there are daily patterns across households in general, with high solar insolation during 8am and 6pm period. The instantaneous demand is higher in the evening than during the night
  - For consumers without solar installation, similar patterns can be observed in the instantaneous demand in general
  - Notably, there are minor differences in patterns across households
- **Season pattern:**
  - For solar-connected customer, the maximum solar insolation occurs in summer in general.
  - For customer without solar installation, the demand may vary according to individual customers' preference of energy use

What customer insights are provided by the analysis (2)

- The instantaneous demand for all customers presents a clear daily seasonality, but no obvious trend is observed

- The stationarity test (p value < 0.05) indicating the instantaneous demand time series data is stationary

- The forecasting model using traditional machine learning is not decently performed, indicating more quality features needed to be included

- The forecasting model using time series theory (e.g., EWMA) achieves the best performance so far

Future improvements for an in-depth analysis to provide more insights

- Data processing: Spark
- Data cleaning: some customer sites have missing data, remove outlier, standardization
- Feature engineering: introducing more meaningful features (e.g., holiday, temperature/humidity, etc., information), first/second order differencing
- Forecasting model: SRIMA, Prophet, LSTM
- In-depth analysis: including demographic data (e.g., income, education, age, family size), postcode level, etc.,

Thank You