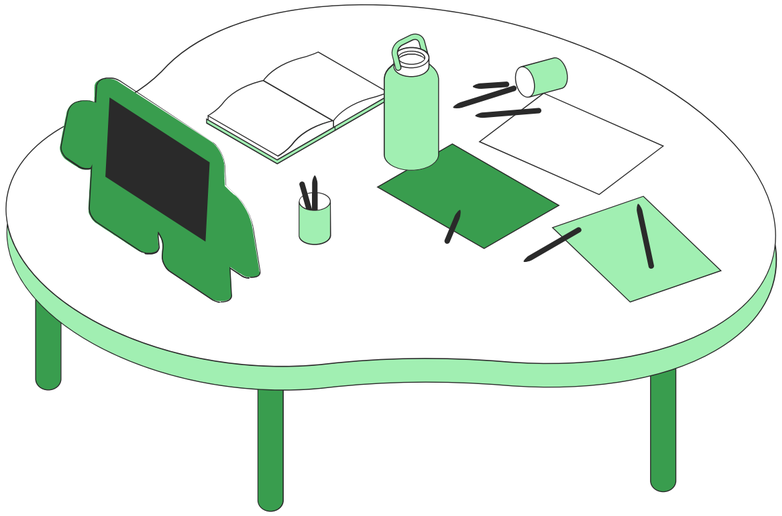
**Final open presentation**

# Imputation Project team

Adrien Lucbert,
Albert Corson,
Jesús Martínez,
Julien van der Niet,
Michael Weij,
Ramon van der Elst

# Table of content



# Terminology

## Imputation

Replacing missing data with a substitute value

## Time-series data

Data indexed by a timestamp

## Building Management System (BMS)

System that controls and monitors building appliances such as a thermostat or heatpump

## Data measurement scale

Measurement refers to the assignment of numbers e.g. Interval, Ratio

## Scenario

Combination of data measurement scale and gap sizes

## Trends

Behavior of patterns In data.

# What is the project about?

## Applied Data Science and imputation

### Imputation

Impute Building Management System time-series data.

### Test methods

Multiple imputation methods studies and tested.

### Write guideline

What imputation methods should be used when?

# Datasets used

## BMS time-series

- Source Net Zero Energy Building from FactoryZero
- Target dataset for research
- 5-minutes measurement interval
- Row count: 105096

## Meteorogical time-series

- Source Royal Netherlands Meteorological Institute (KNMI)
- Contains similar data to BMS time-series
- Hourly measurement interval
- Row count: 17545

# Selected columns

Data selected for measurement scales

| Column | Dataset | Scale |
|---|---|---|
| Flow_temp | BMS | Interval |
| Temperature | KNMI | Interval |
| OP_mode | BMS | Nominal |
| Power | BMS | Ratio |
| CO2 level | BMS | Ratio |
| Solar Radiation | KNMI | Ratio |
| Humidity | KNMI | Ratio |

# The Testing Method

### Loading data

Load data without gaps

### Creating gaps

Create reproducible gaps of different  sizes

### Running imputations

Hot-swappable imputation methods

### Evaluating results

Wide range of evaluation criteria

# Imputation methods

- **Wide scope of methods**
- **From previous literature**

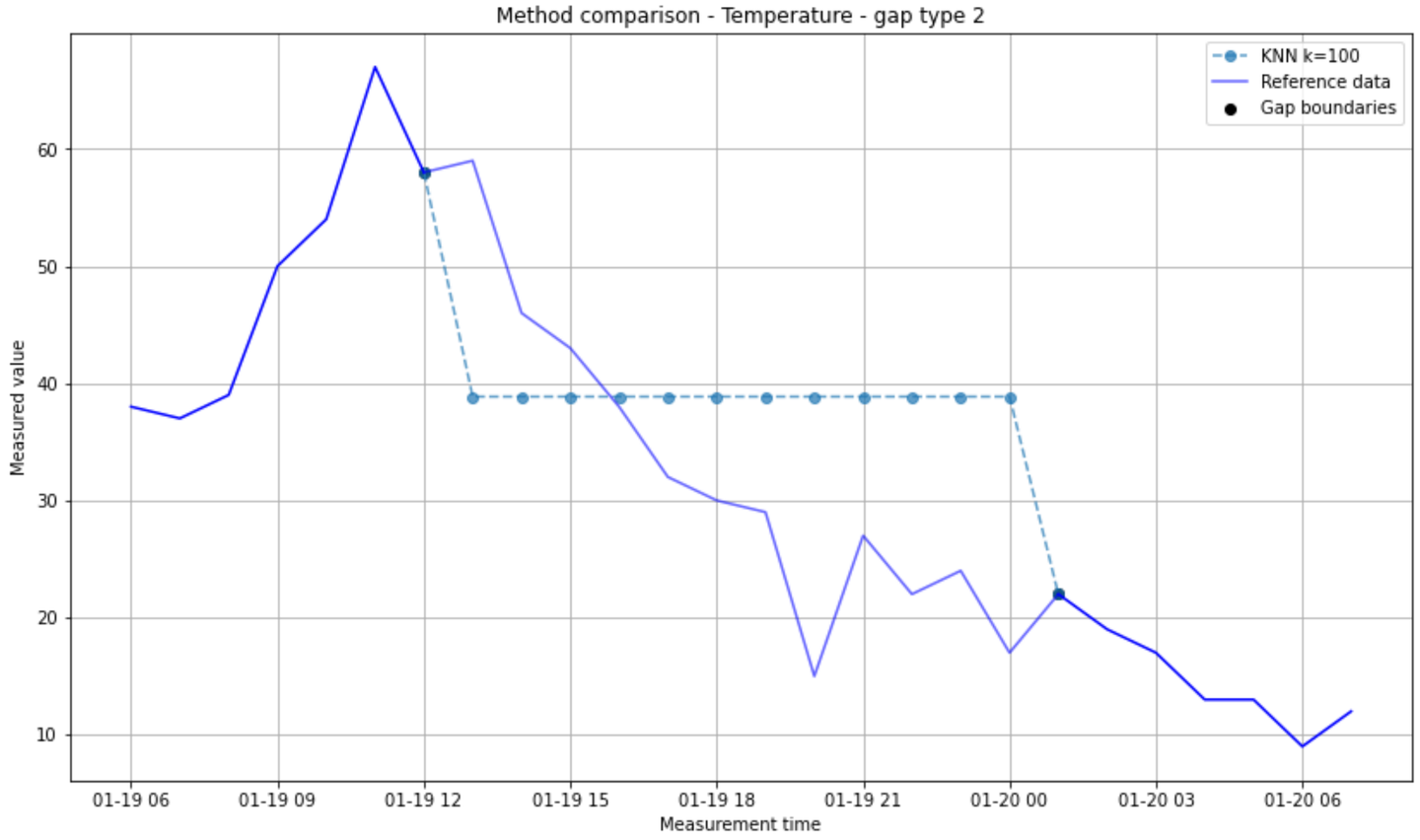
- **K-Nearest Neighbour regression (KNN)**
- **Last Observation Carried Forward (LOCF)**
- **Recurrent Neural Networks (RNN)**
- **Hot Deck (HD)**

# K-NN regression

**The K-values tested are: 1,5,10,15,20,100**
Compared by Root Mean Squared Error

**Best results:**

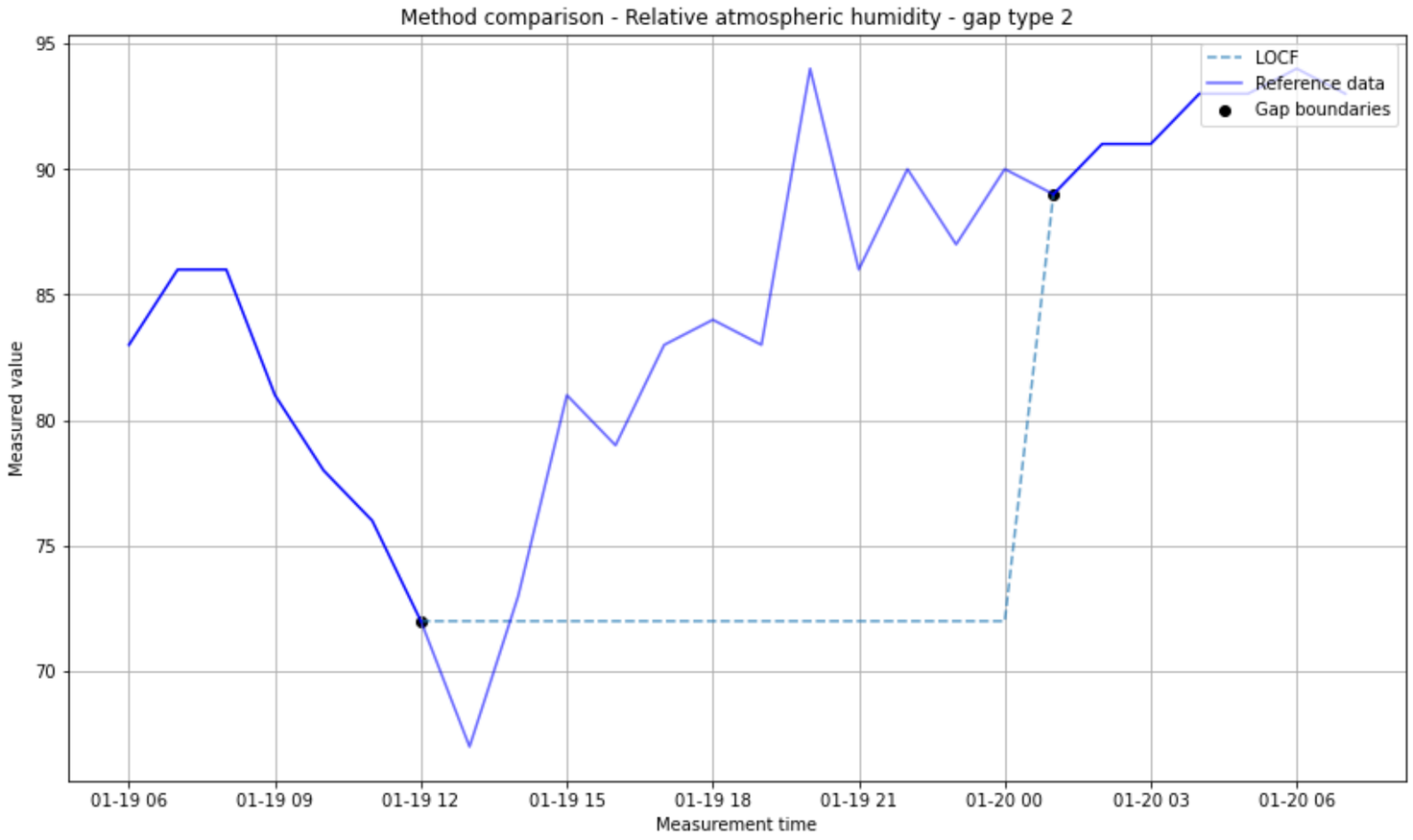
Gap size 1: KNN = 5
Gap size 2,3,4,5: KNN = 100



# LOCF

**Last Observation Carried Forward**

Use the last valid measurement to fill a gap



# Hot Deck

## Use data from donors

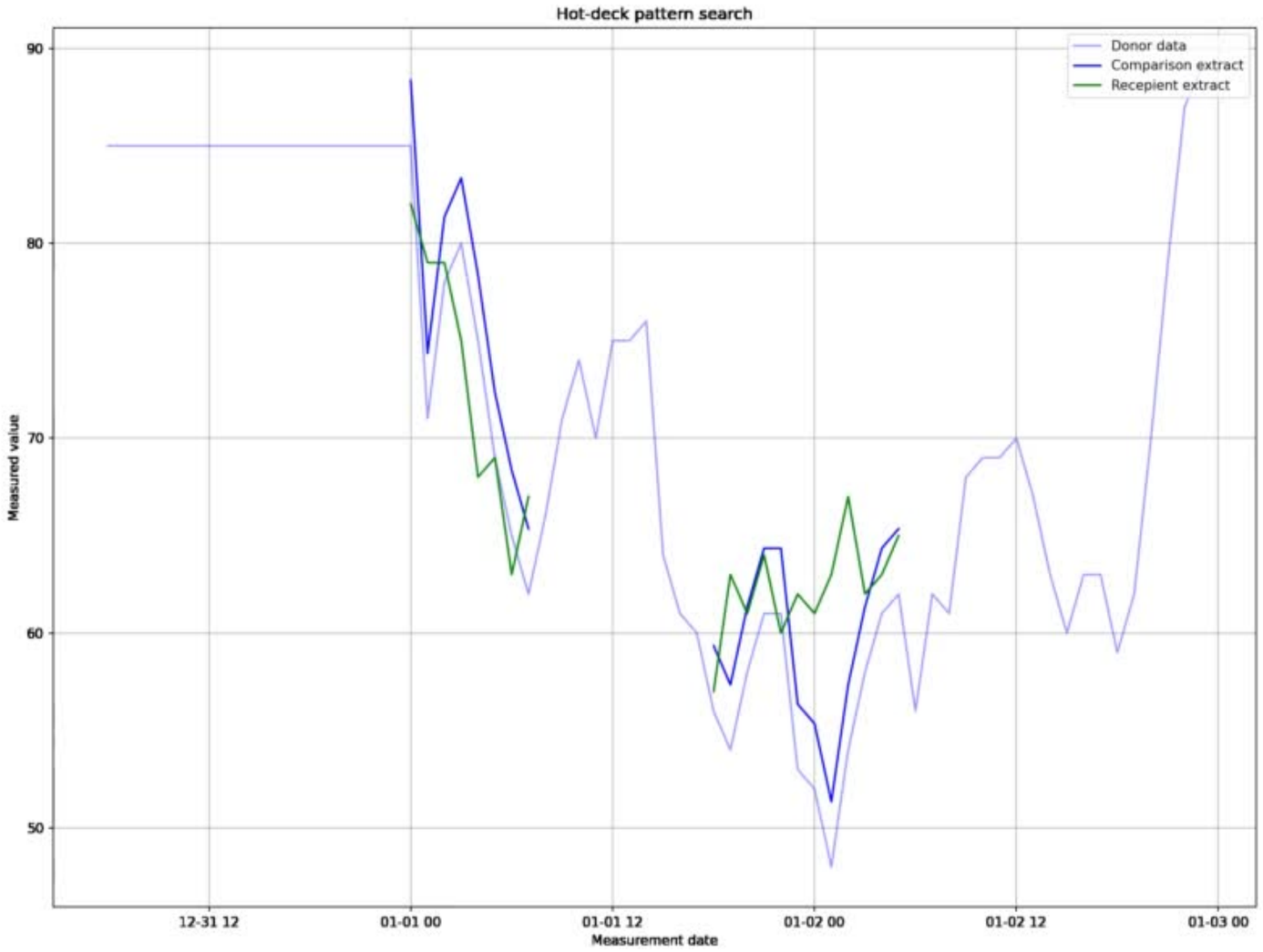
Import data from similar units

## Donor selection

Pattern search based selection

## Reliance on donor quality

Heavily dependant on the quality of the donors



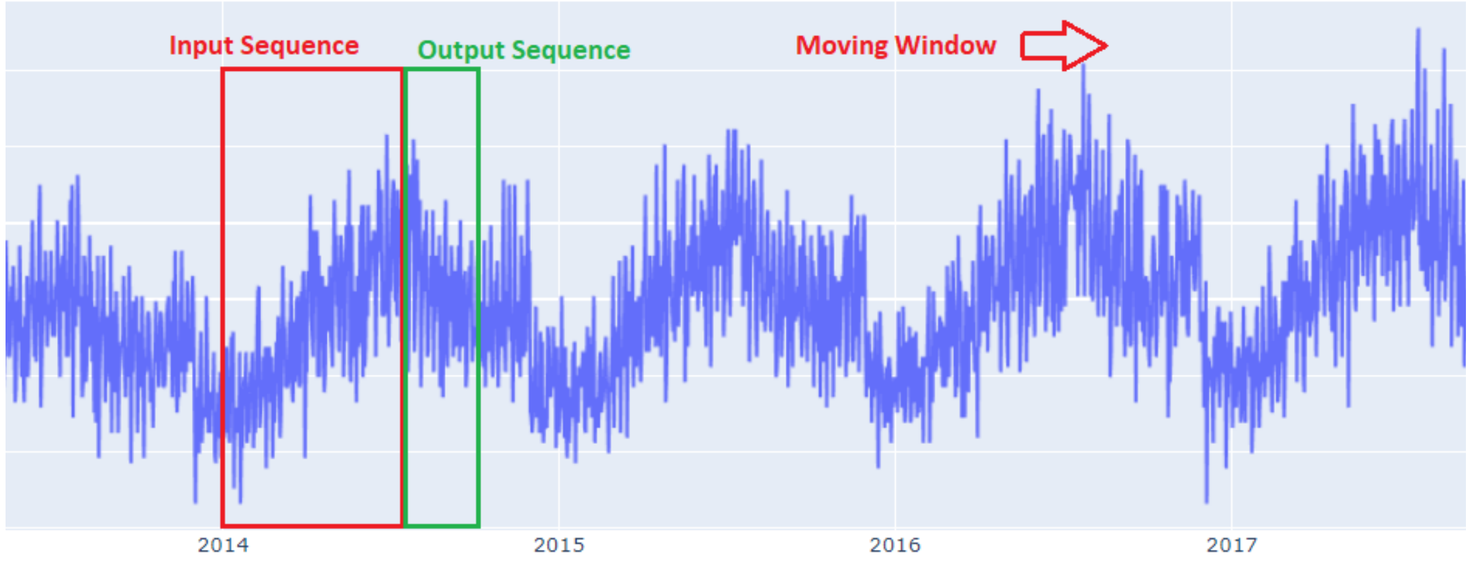
# Recurrent Neural Network

## Sequence as input

Considers a sequence of measurements to predict one or a sequence of the following values

## Fit for time-series

Time-series values often depend on values preceding them, which makes RNN ideal for predicting time-series data



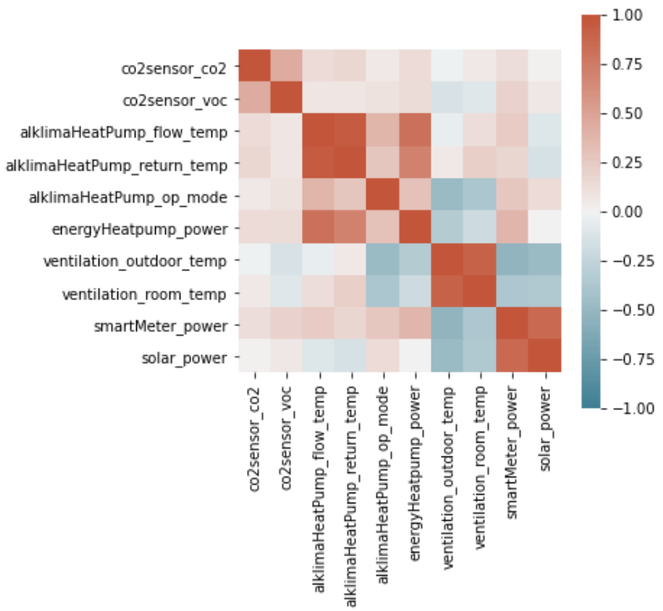
# Recurrent Neural Network

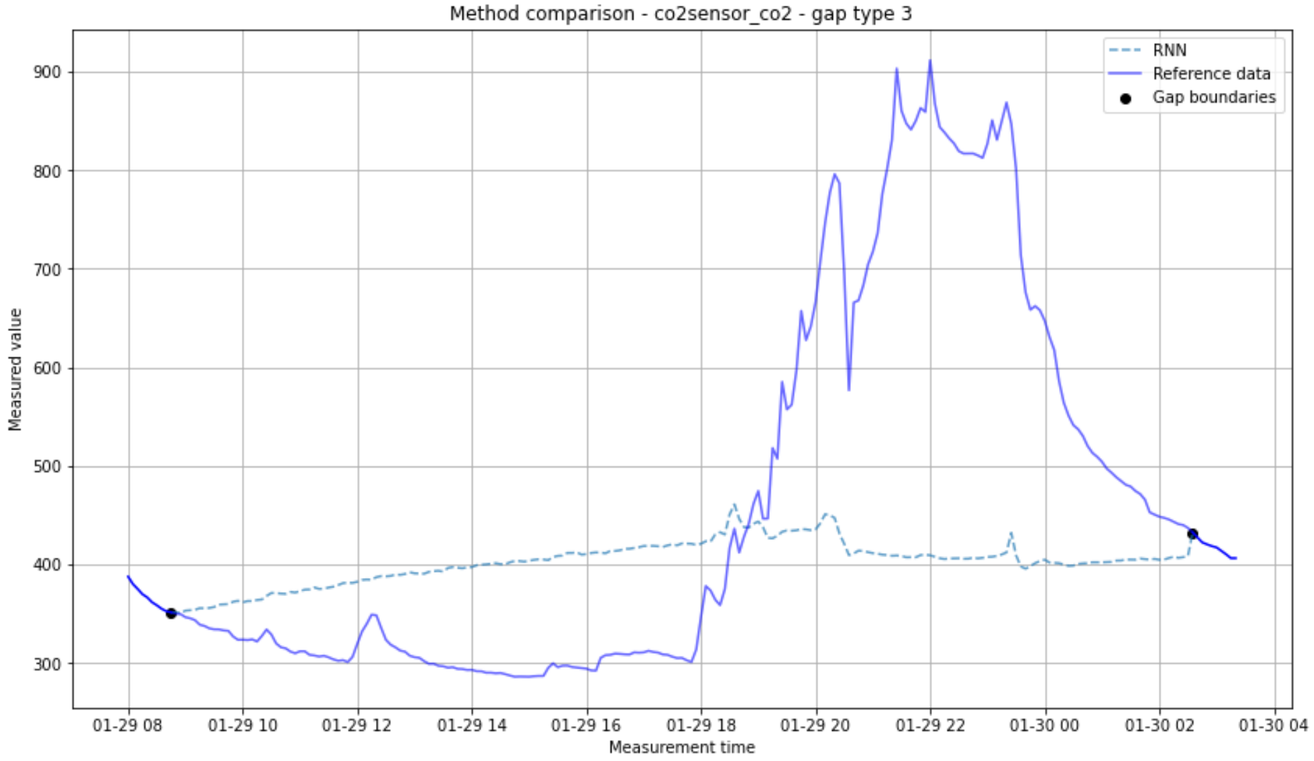
## Makes use of correlation

It always performs much better with highly correlated features

## Poor correlators = poor predictions

Predictions need correlators, but poor correlators will still result in poor predictions



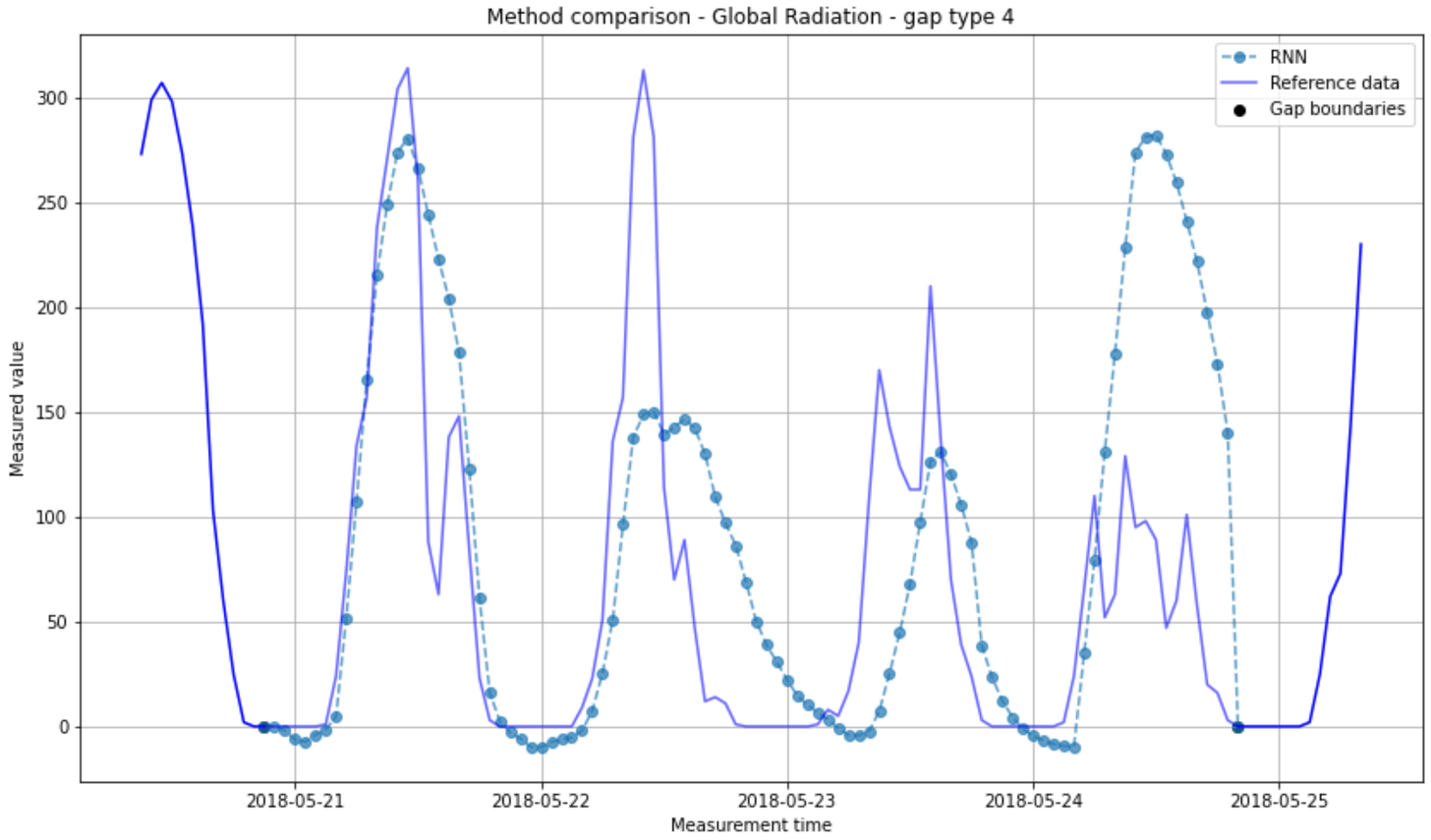
# Recurrent Neural Network

**Great with interval data**

We noticed it performs especially well with interval data

**Reliance on correlators**

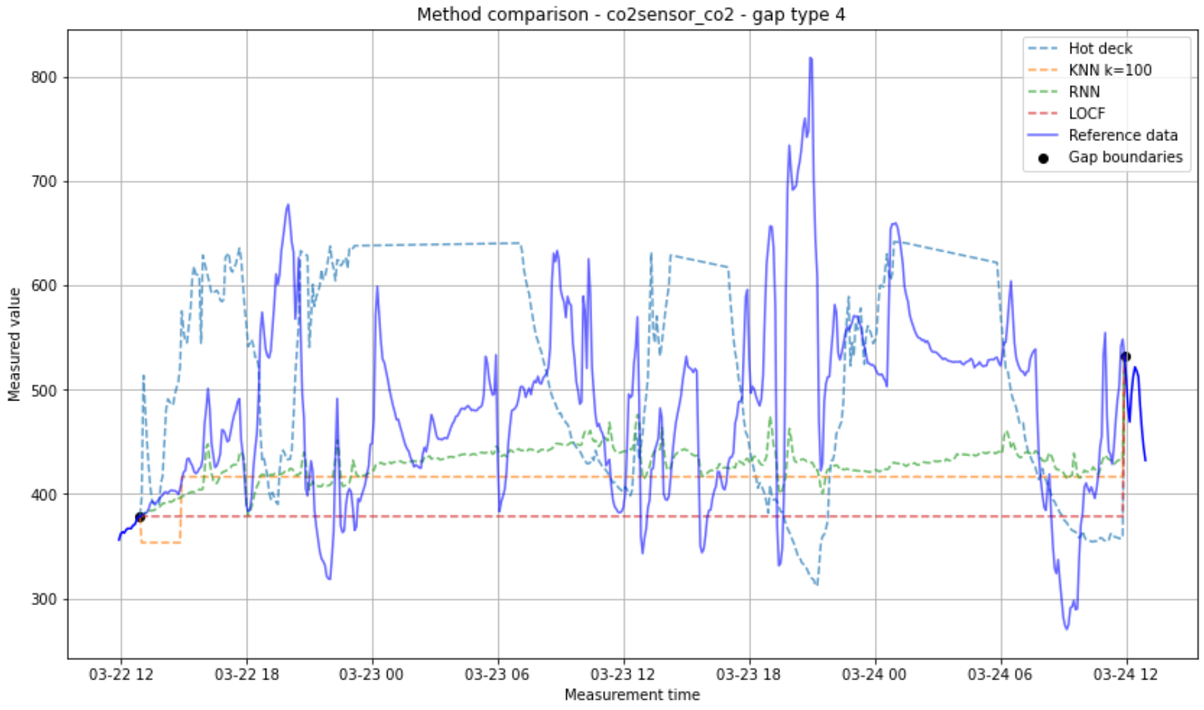
Consistent correlators = great predictions



# Results

## Smaller conclusions based on results

- Imputation, as evaluated by RMSE, is poor
- Hot Deck performs better on KNMI
- RMSE and VE results don't always align
- No consistent link between performance and difference in Kurtosis & Skewness



# Conclusion

## Resulting guideline

- Guideline in both VE and RMSE
- No single best method for all gap sizes and scales
- Performance on gap depends on scale

| | Gap size 1 | Gap size 2 | Gap size 3 | Gap size 4 | Gap size 5 |
|---|---|---|---|---|---|
| **Nominal** | HD | HD | HD | HD | HD |
| **Ratio** | HD | HD | HD | HD | HD |
| **Interval** | RNN | RNN | RNN | RNN | RNN |

# Future work

**FOCUS:**

**-** evaluating imputation with metrics based on the error

**+** evaluating on the impact of forecasting using imputed data

**Sequence-to-sequence RNN**

Instead of sequence-to-value RNN in order to remove potential bias of imputation using its own imputed values.


Do you have any questions?