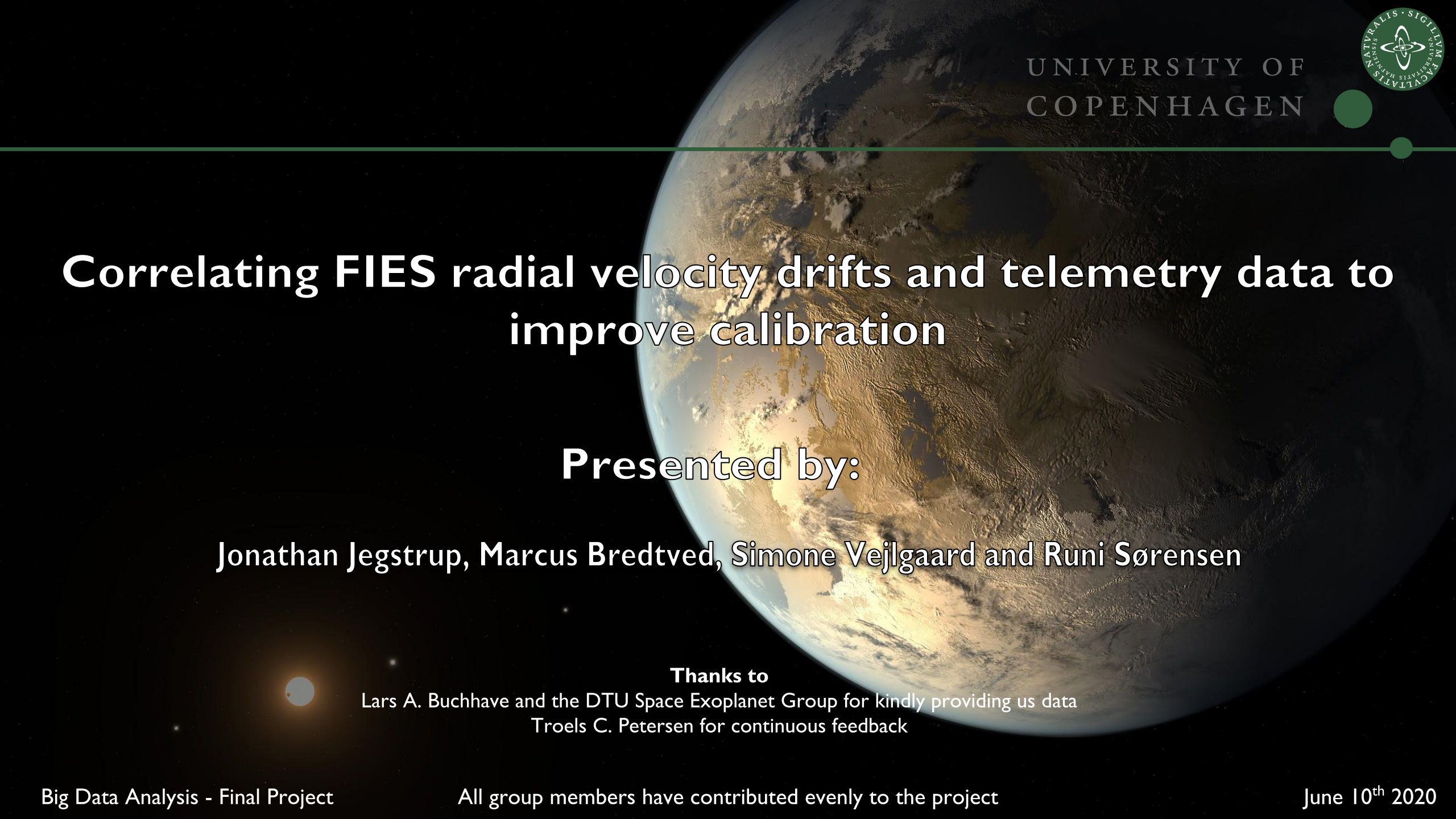


UNIVERSITY OF COPENHAGEN

Correlating FIES radial velocity drifts and telemetry data to improve calibration

Presented by:

Jonathan Jegstrup, Marcus Bredtved, Simone Vejlgaard and Runi Sørensen

Thanks to
Lars A. Buchhave and the DTU Space Exoplanet Group for kindly providing us data
Troels C. Petersen for continuous feedback

Big Data Analysis - Final Project

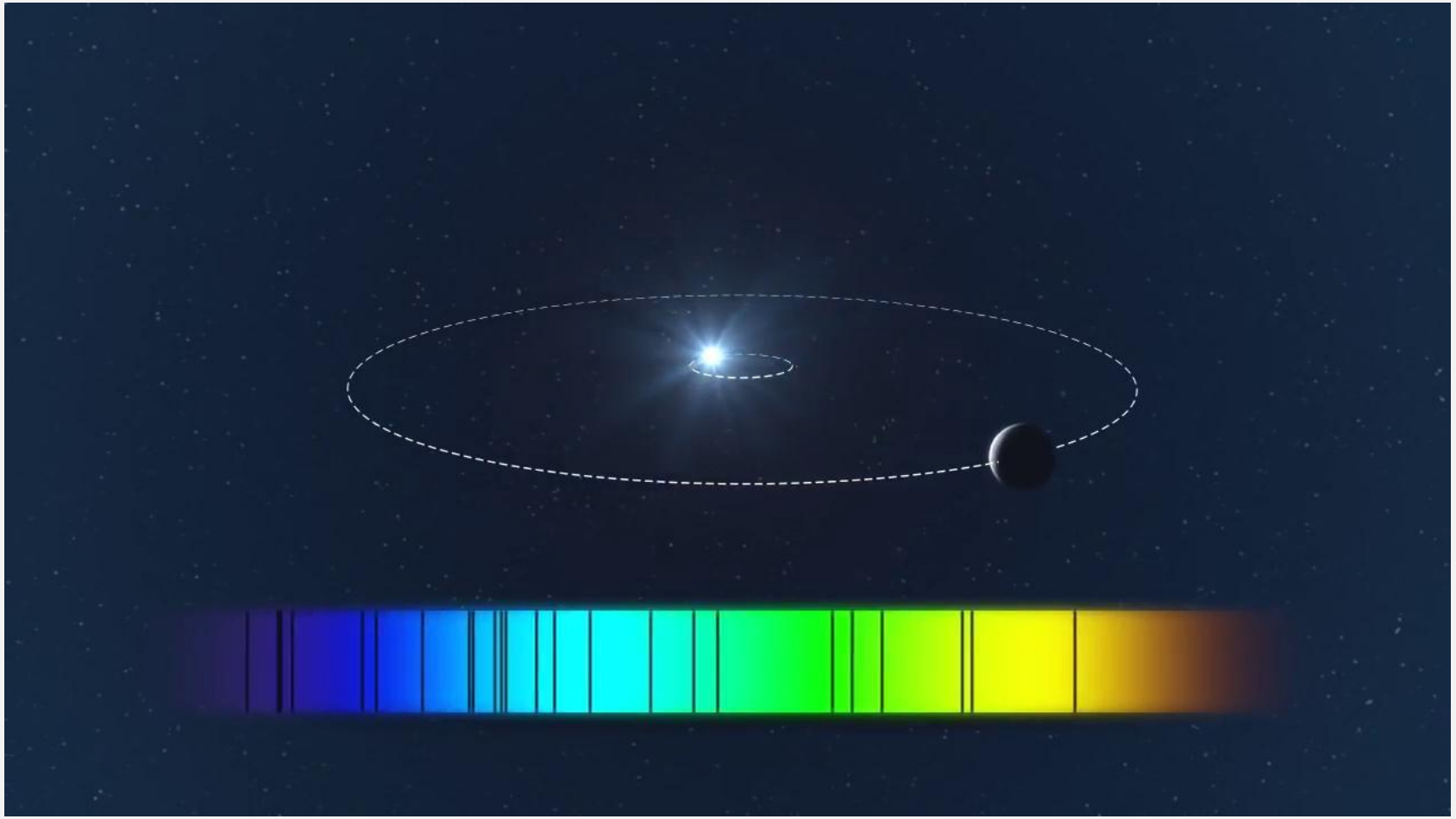
All group members have contributed evenly to the project

June 10th 2020

Introduction and motivation

- Data from exoplanet research group at DTU Space
- ‘Going in blind’ - Uncertain ML applicability, but scientific merit!
- Science case: Improved calibration of radial velocity drift due to instrument drift will increase FIES (spectrograph) ability to provide confirmations and mass measurements for exoplanet candidates discovered by TESS (space observatory) around bright, nearby stars - and help the search for Earth-like exoplanets.

Quick intro to radial velocity method and the spectrograph

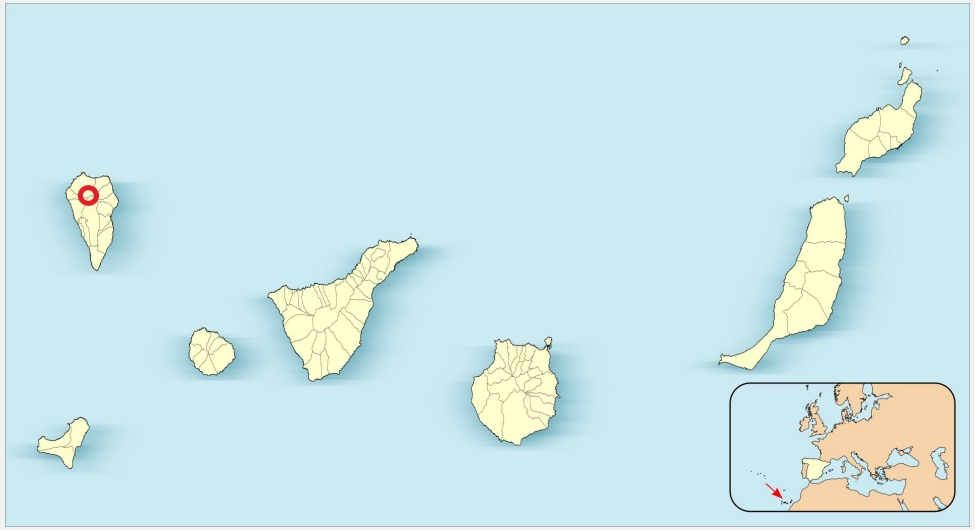


Credit: ESO/L. Calcada

Quick intro to radial velocity method and the spectrograph

Nordic Optic Telescope on the island of La Palma, Canary Islands, Spain

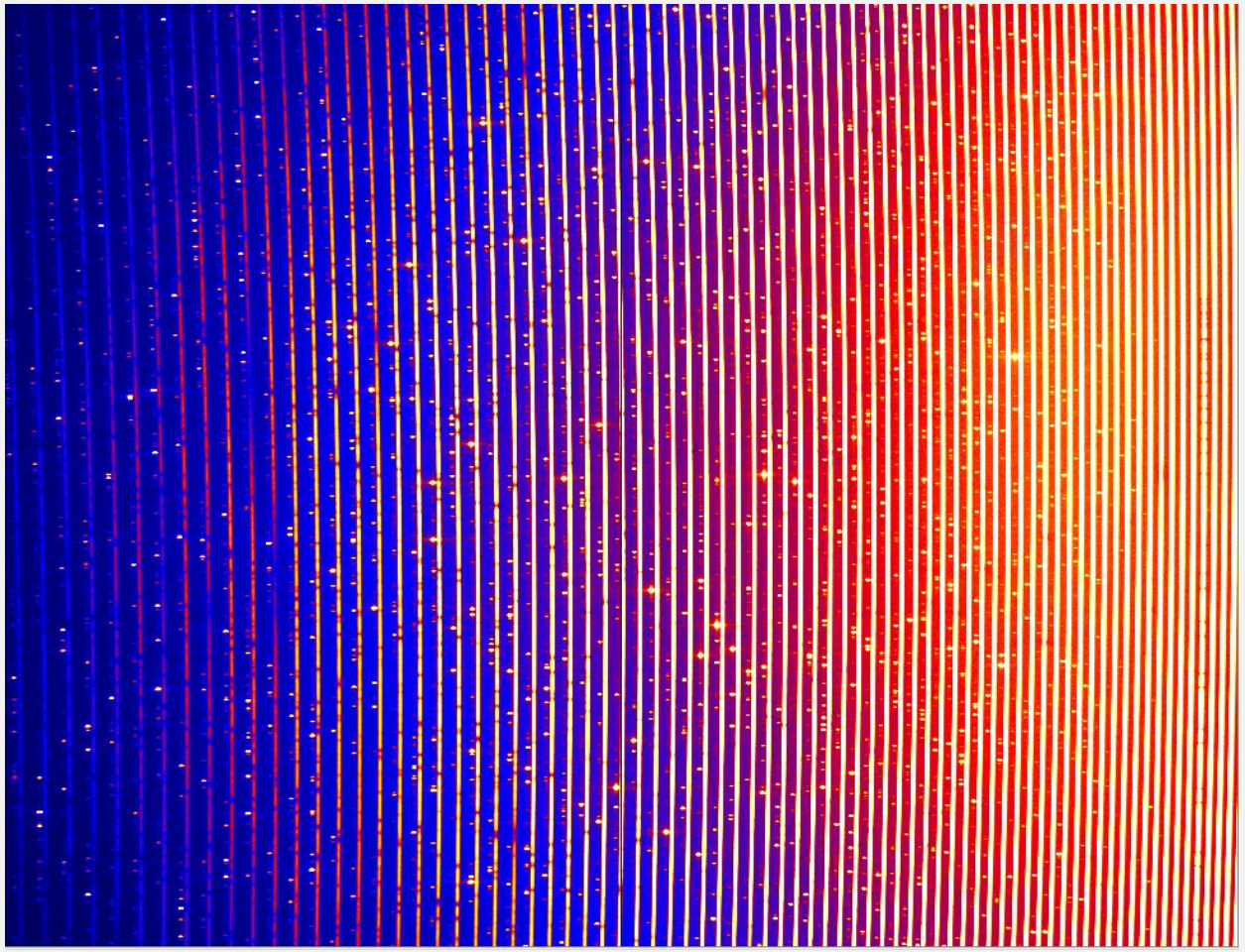
NOT



TESS

Quick intro to radial velocity method and FIES

- Path of light is very complicated - many steps
- Extreme stability is needed for precise RV measurements
  - 1/1000 of a CCD pixel = 1 m/s
- Drift arises from tiny changes due to mechanical or thermal fluctuations - changes the physical properties of the instrument

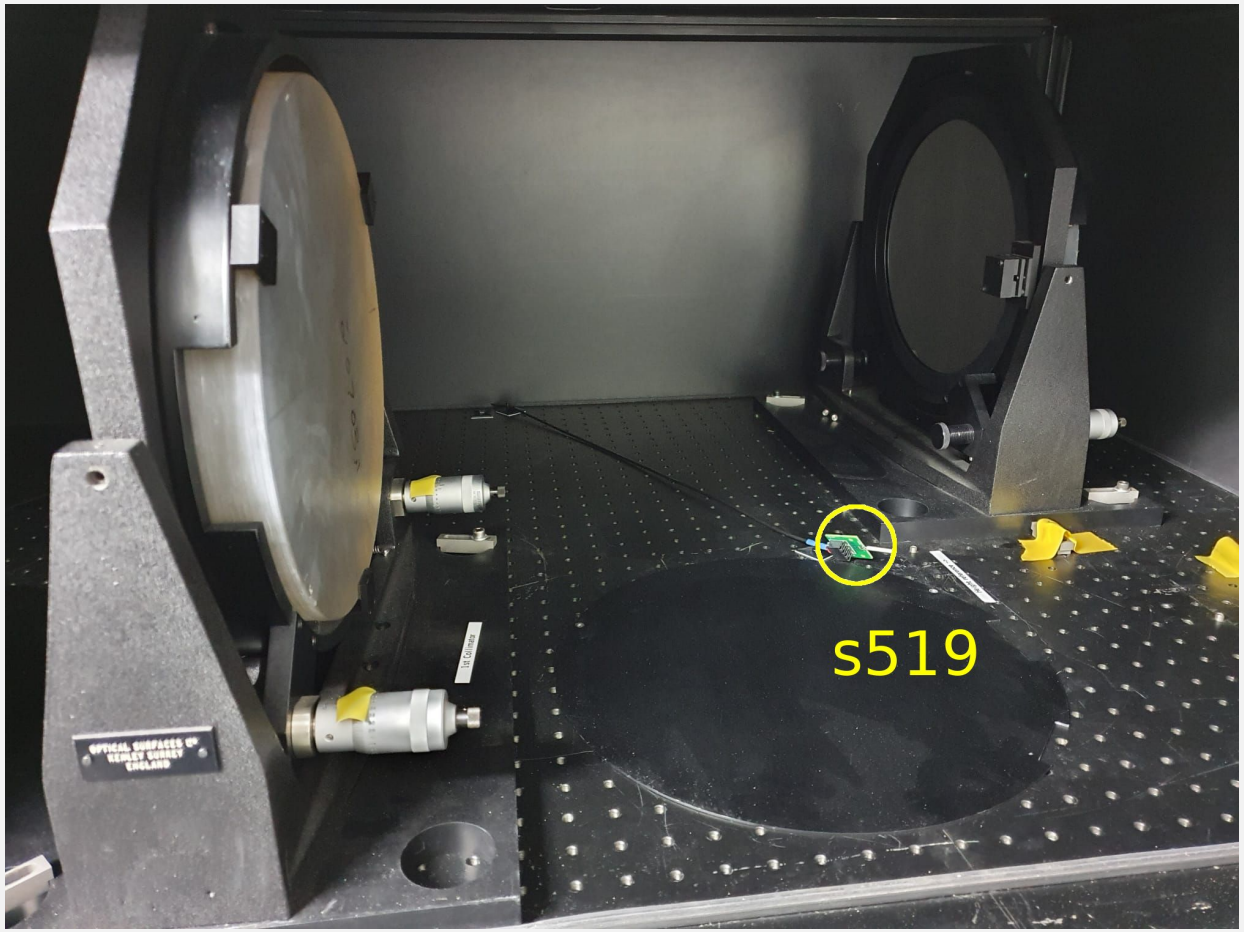


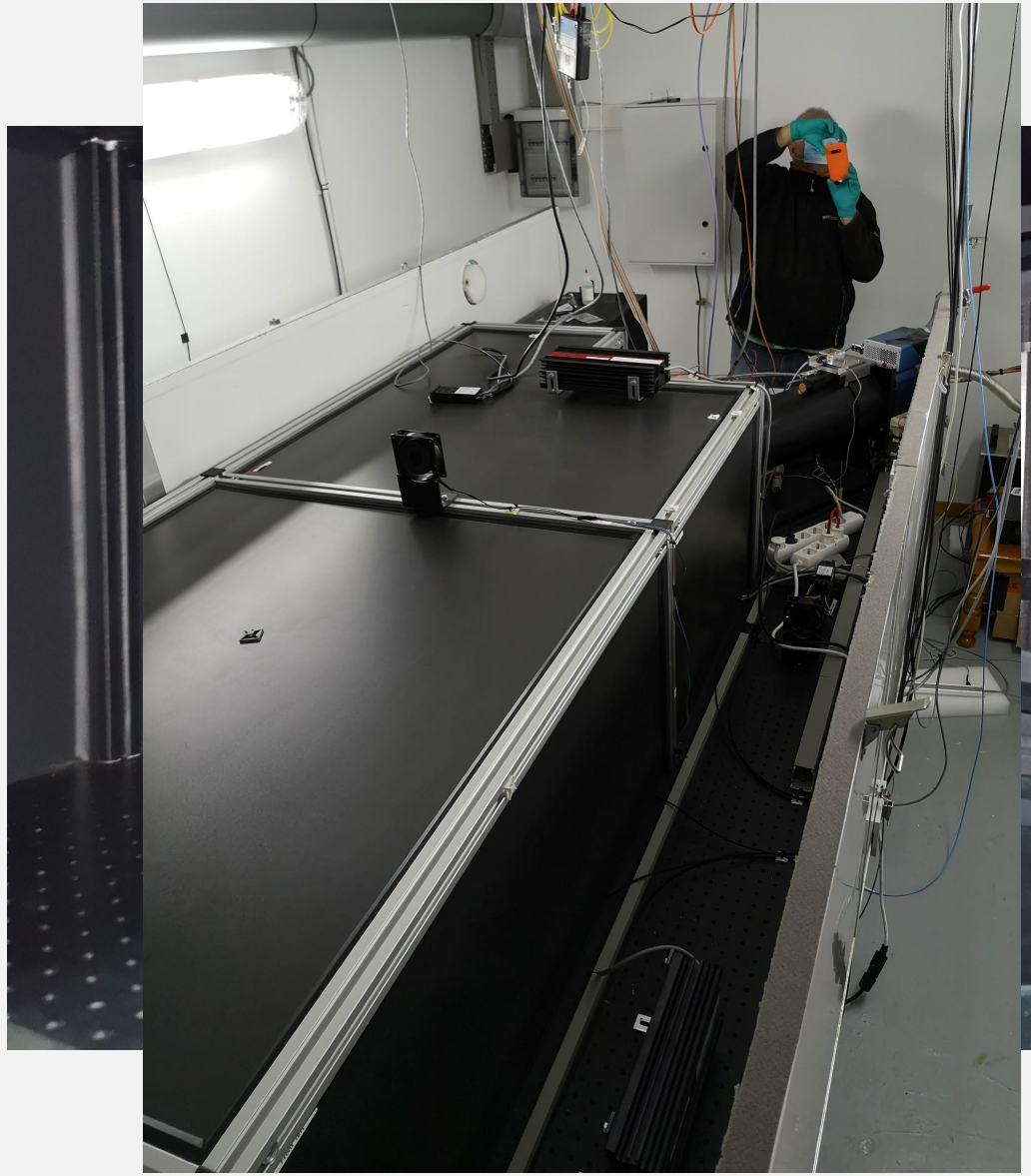
Raw spectrum from FIES

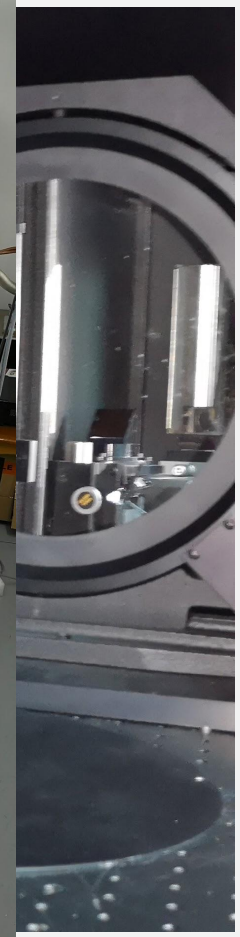
Motivation and goals

- Data of the uncalibrated spectrograph is drifting, likely due to temperature shifts in the spectrograph itself
- Can we understand these drifts based on the available telemetry placed around the instrument?
- Task is to see if a combination of telemetry sensors can train a model that describes the drift - or if we can ascertain that the telemetry itself needs to be improved

Instrumentation - the insides and sensors

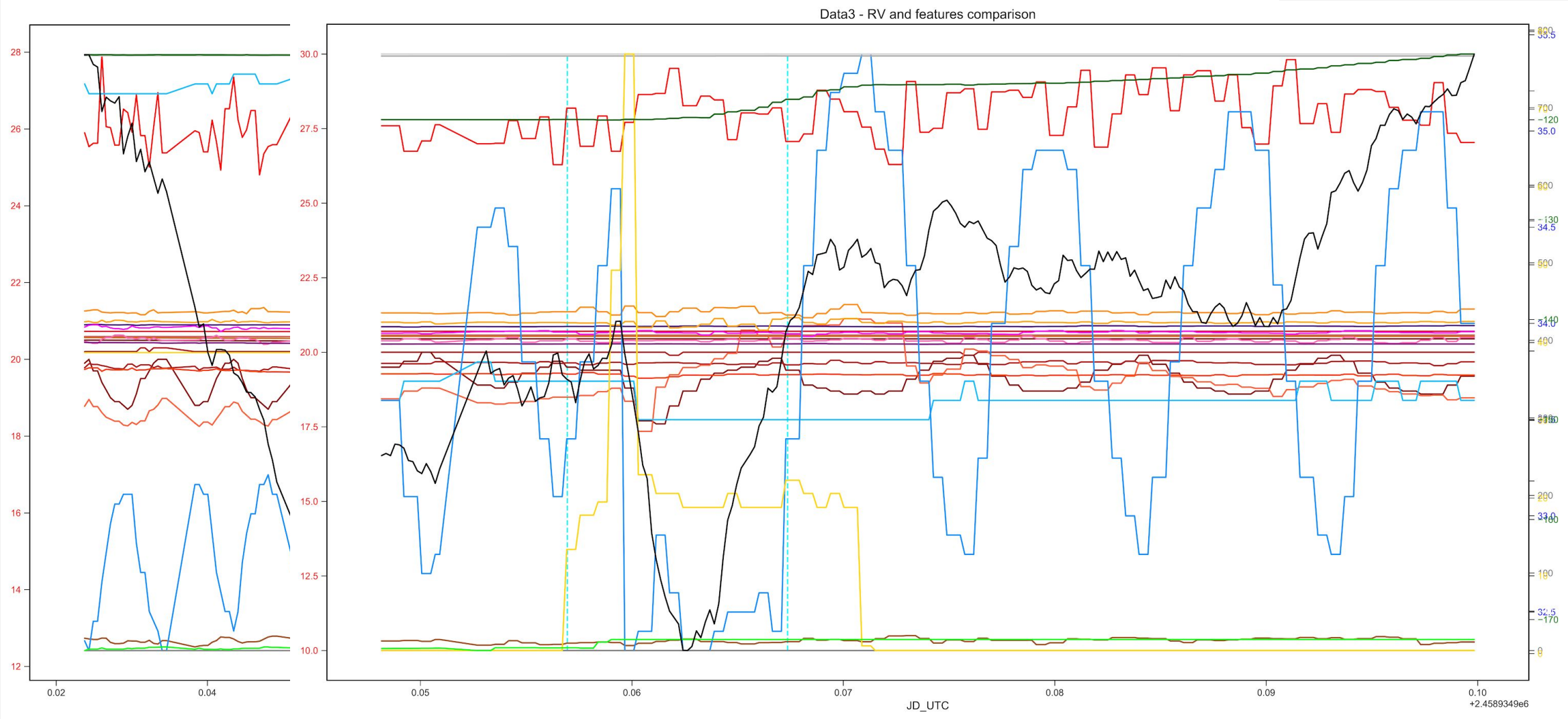




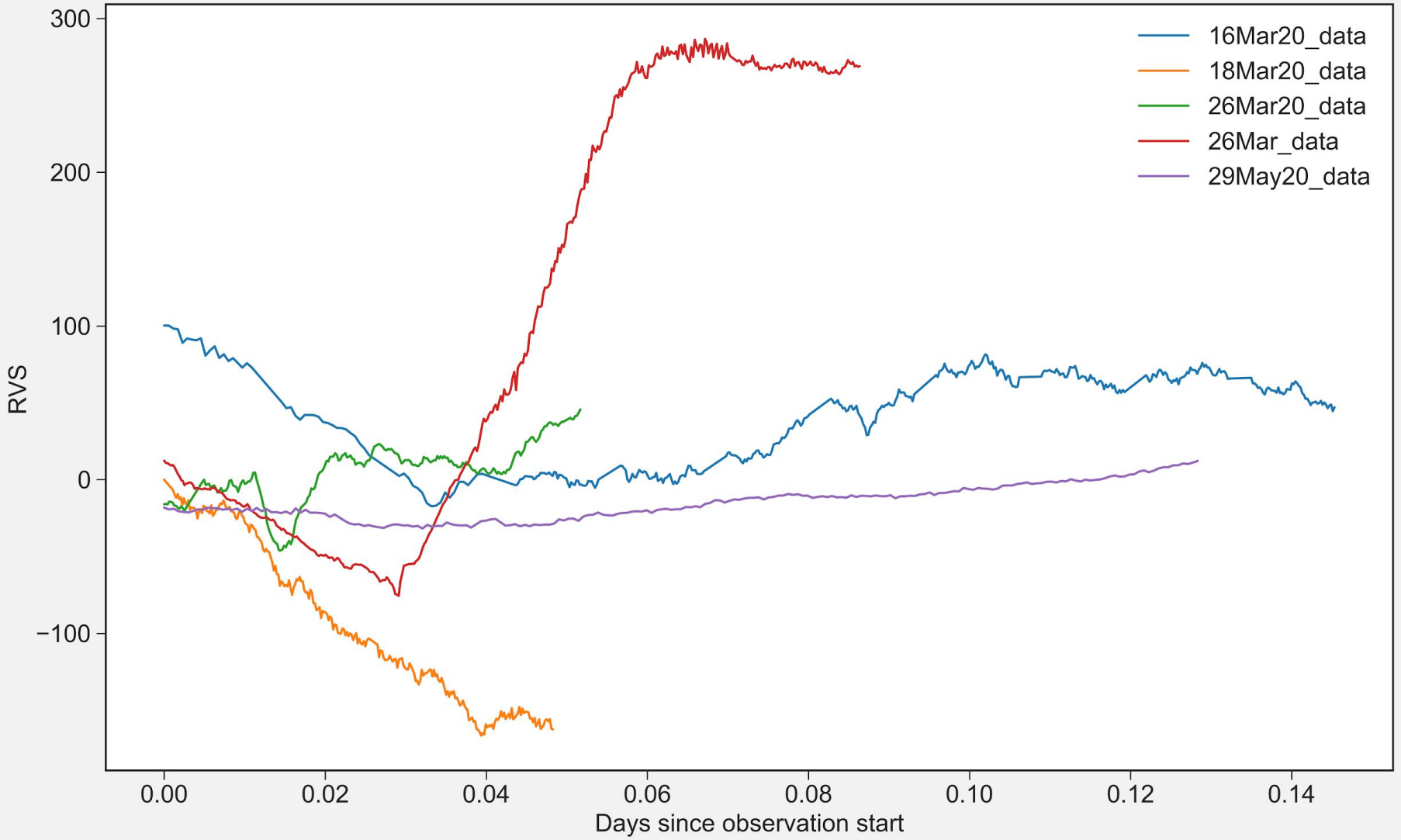


Visualization of the telemetry data

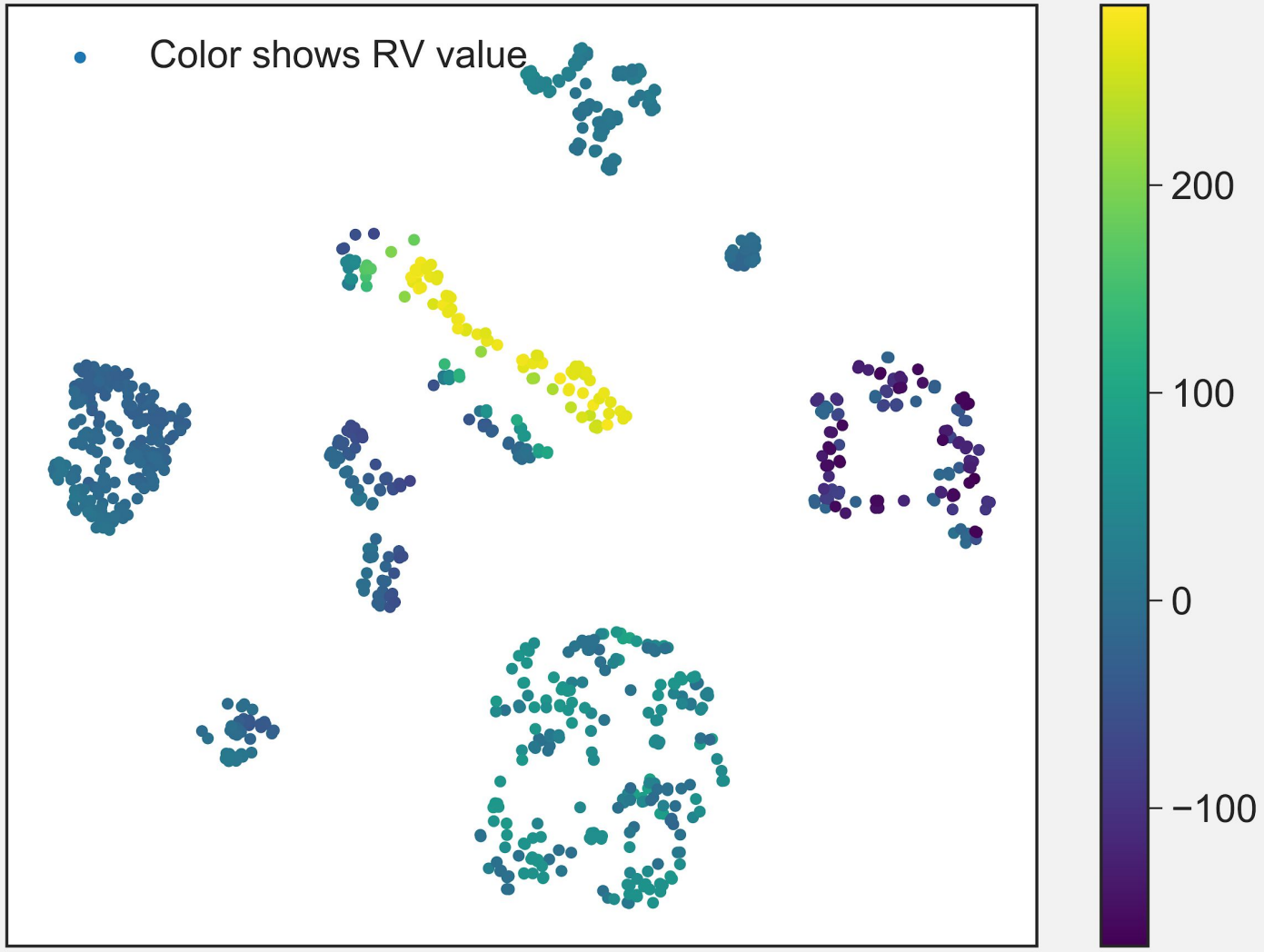
Visualization of the telemetry with RV data



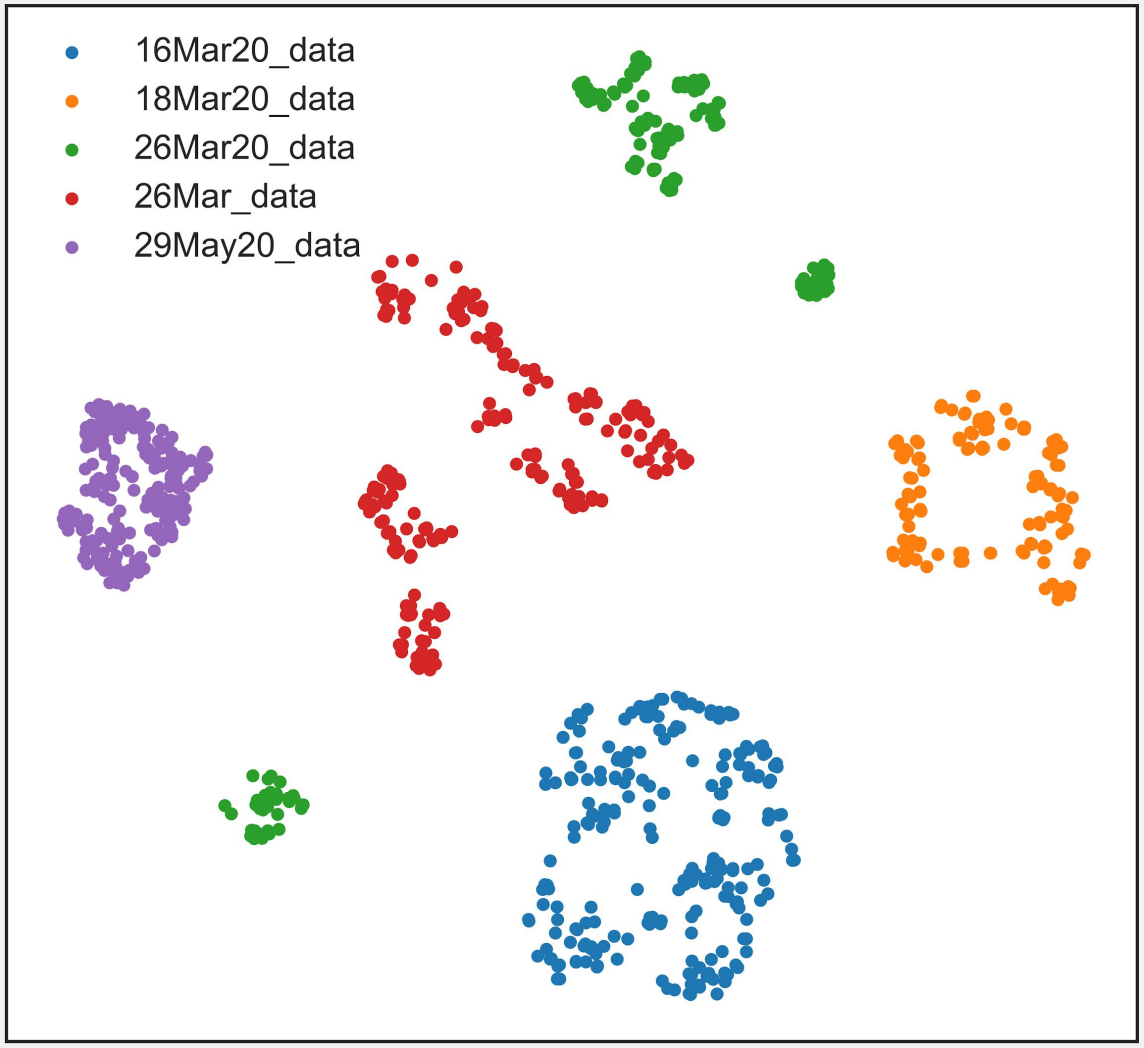
The 5 runs


300
200
100
0
−100
RVS
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Days since observation start
16Mar20_data
18Mar20_data
26Mar20_data
26Mar_data
29May20_data

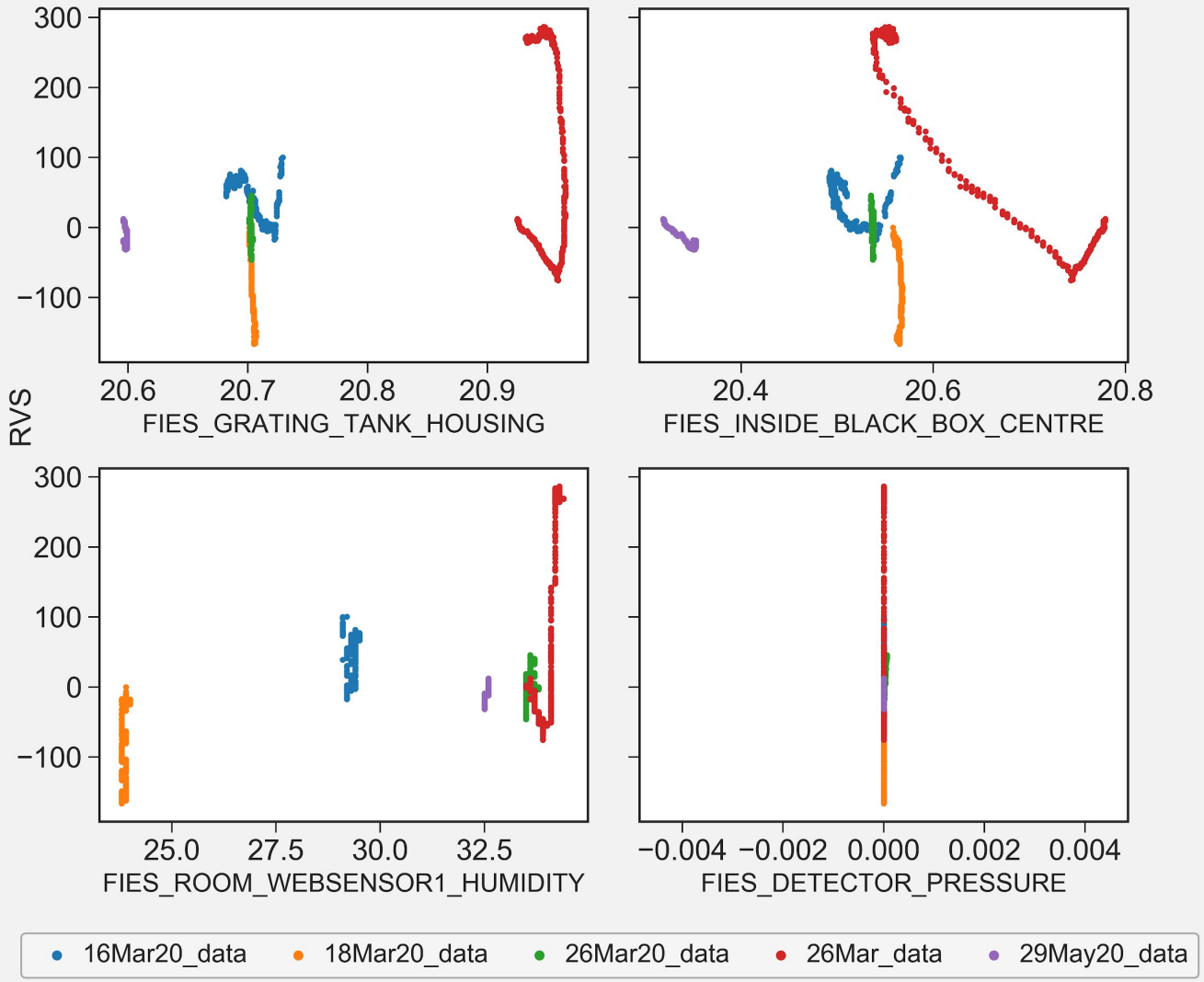
t-SNE dimensionality reduction



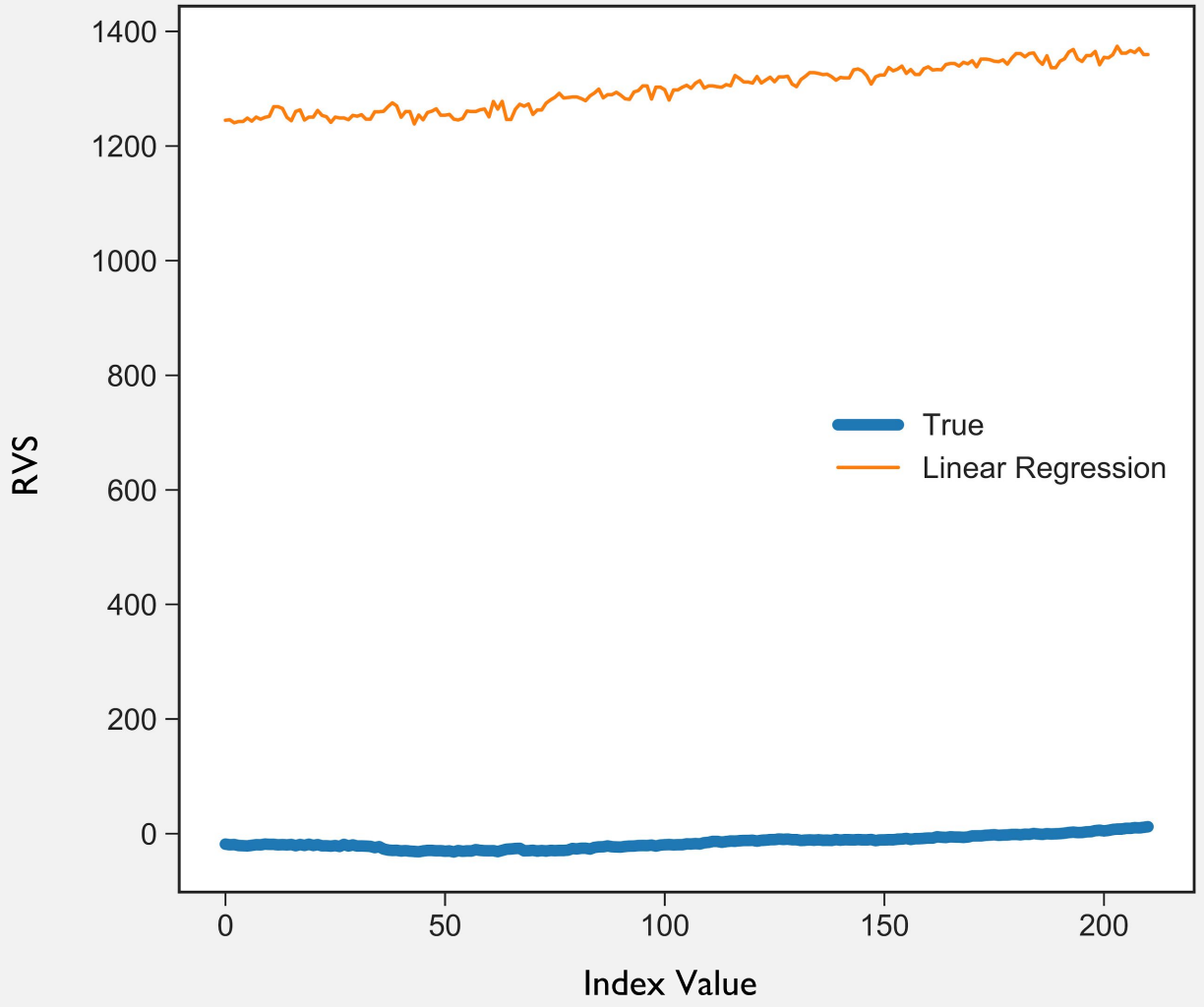
t-SNE dimensionality reduction



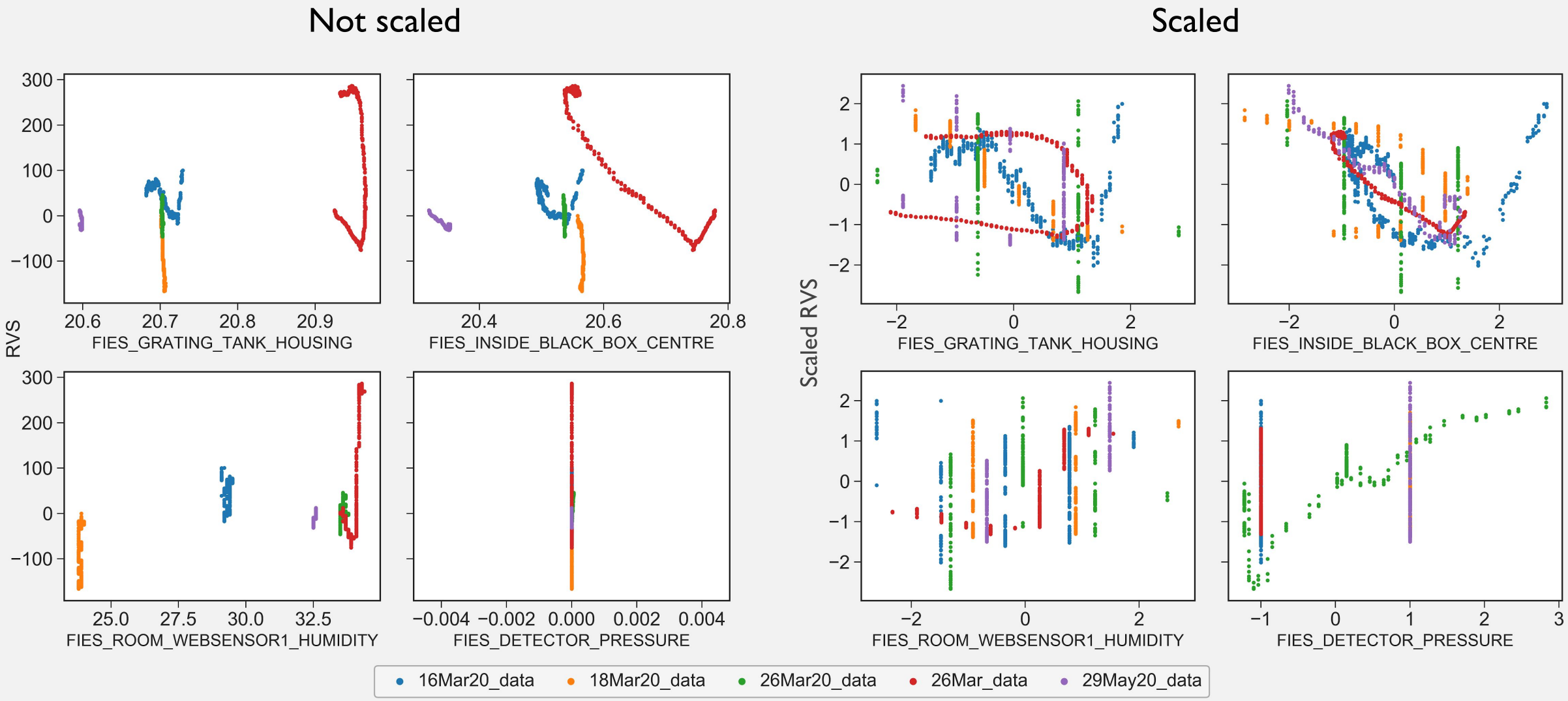
Parameter correlation to RVS



Simple linear regression - Test on timeseries 5

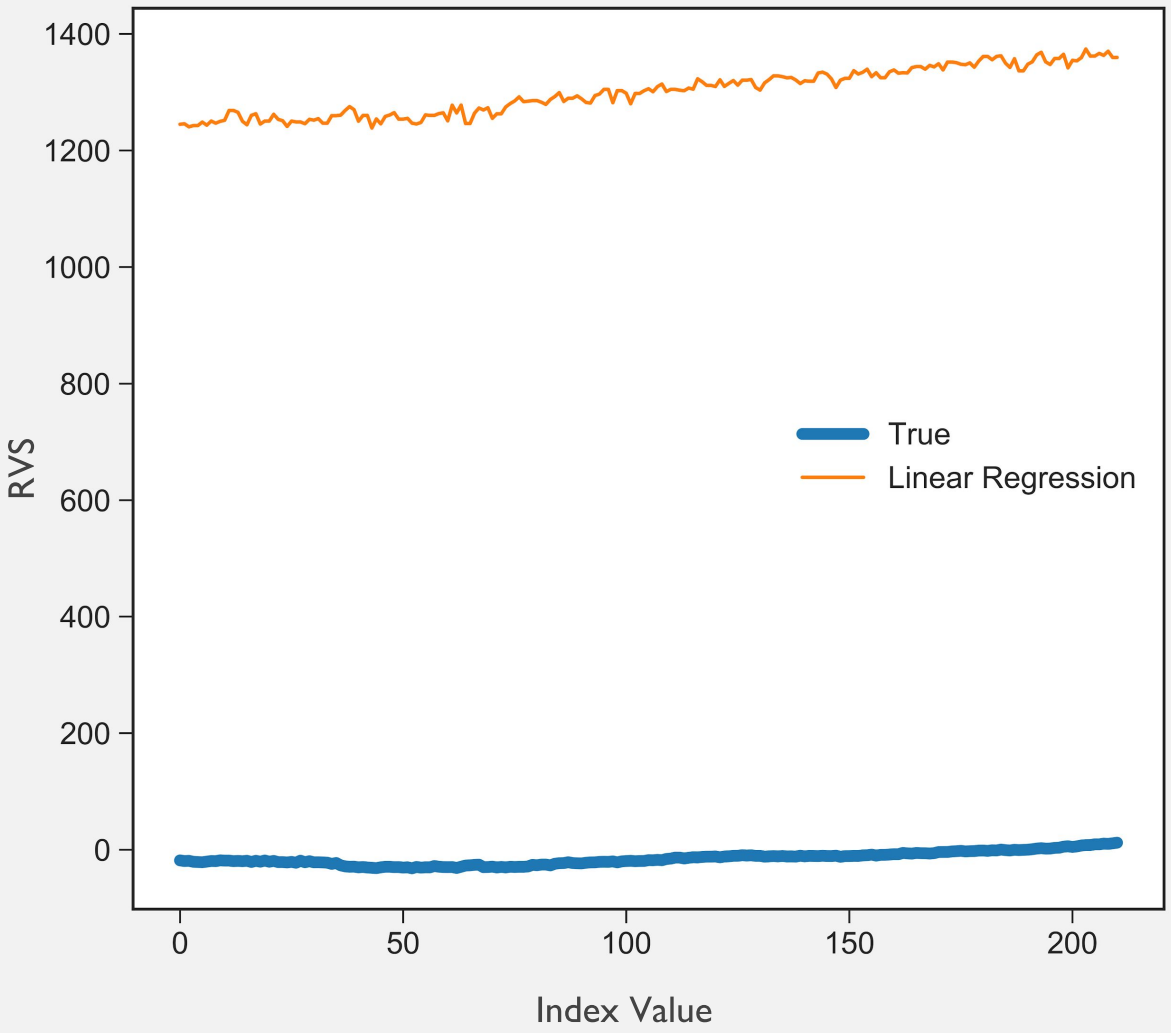


Parameter correlation to RVS

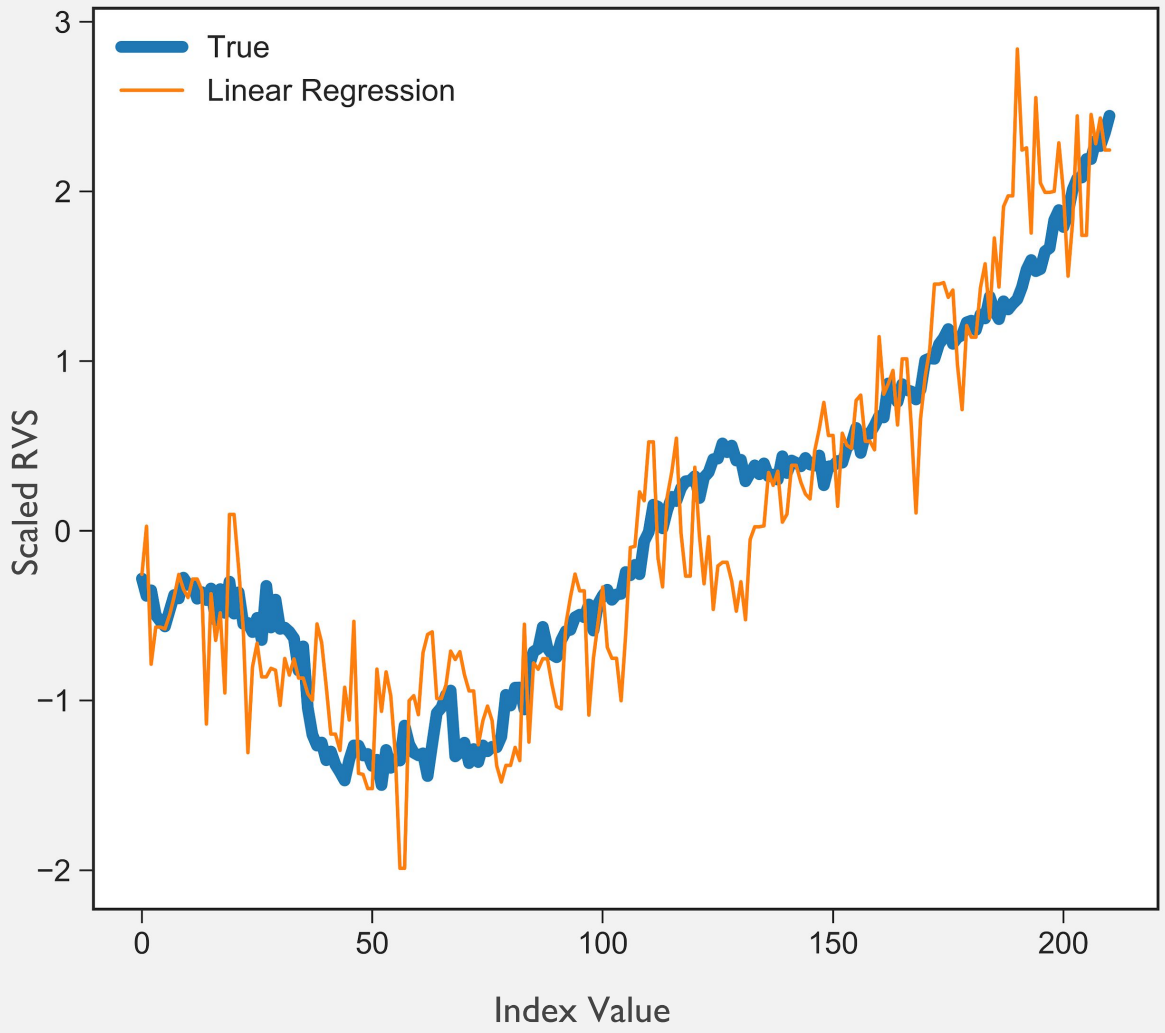


Simple linear regression - Test on timeseries 5

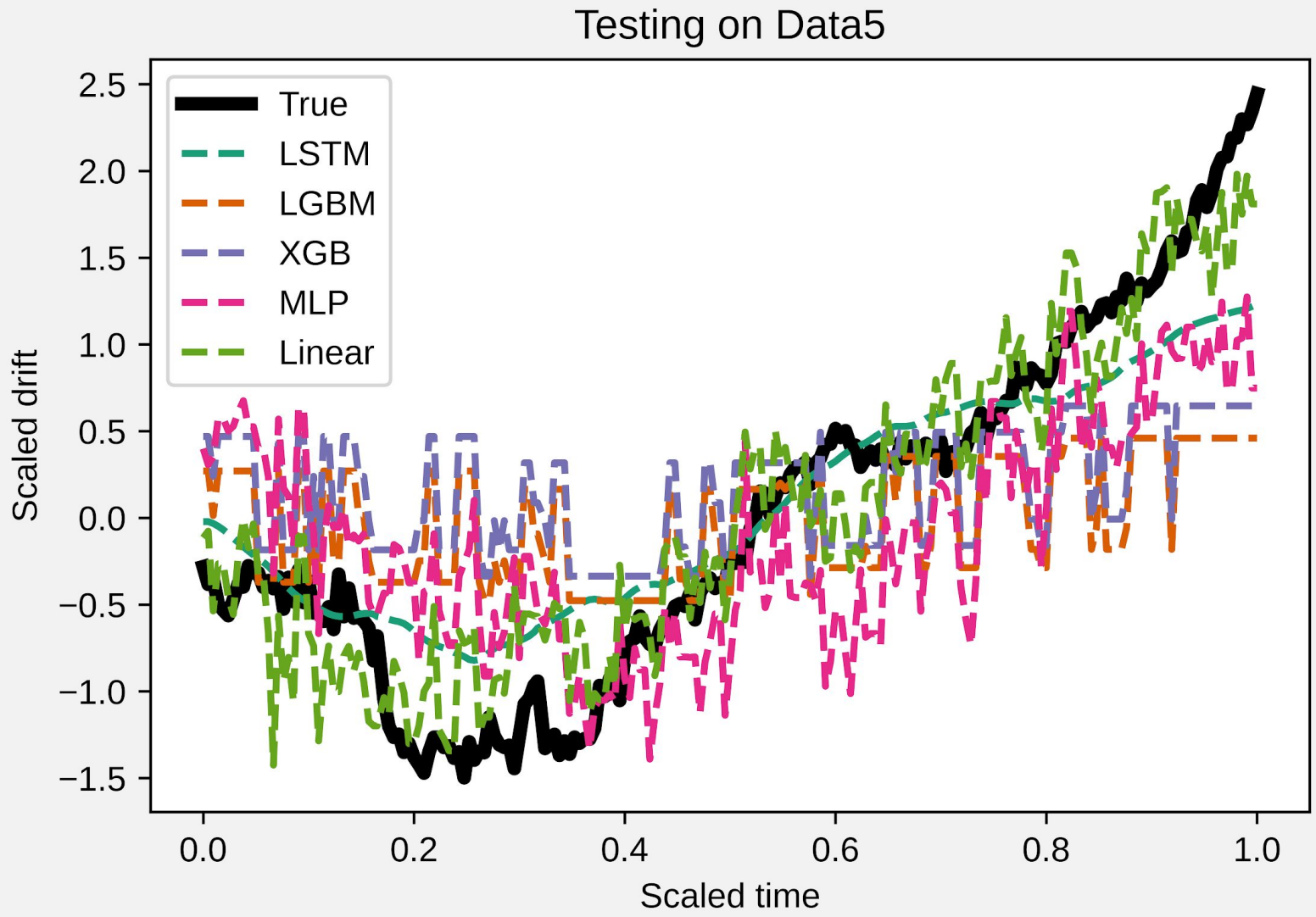
Not scaled



Scaled

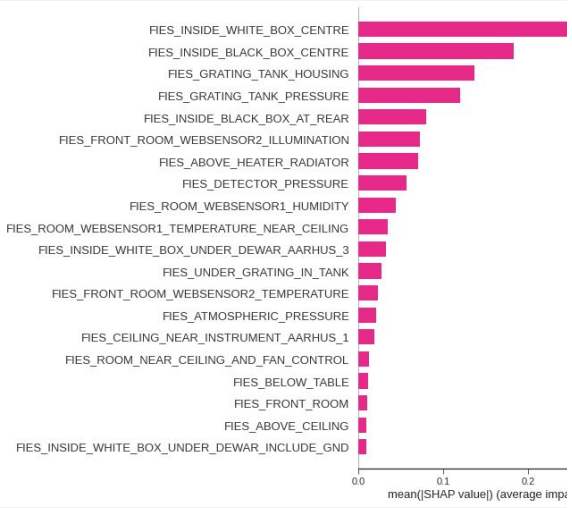


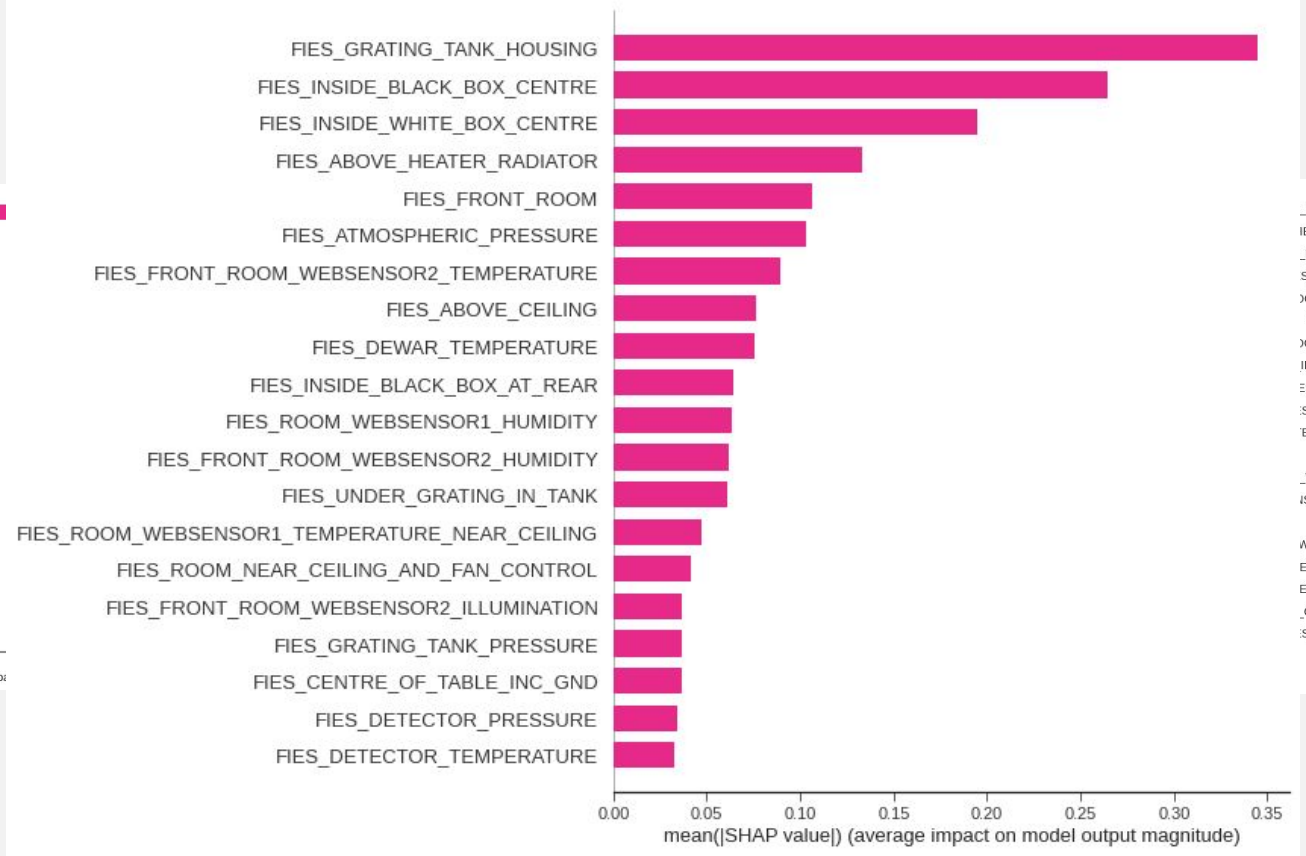
Predictions



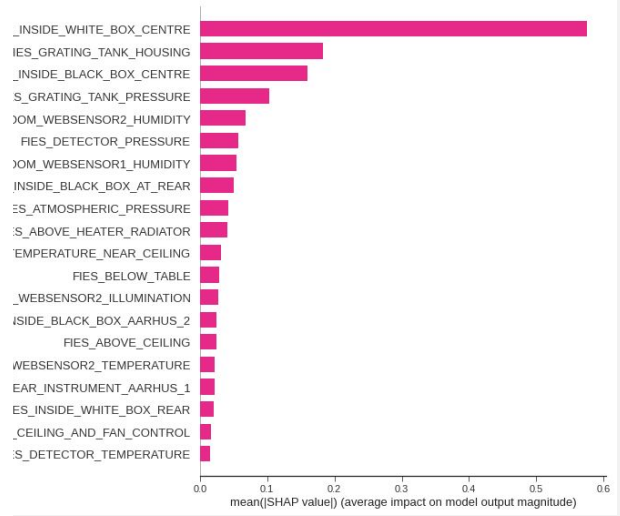
SHAP-values

XG Boost





Light GBM



Coherent with expectations?

Our top features:

1. FIES_INSIDE_GRATING_TANK
2. FIES_INSIDE_BLACK_BOX_CENTRE
3. FIES_INSIDE_WHITE_BOX_CENTRE
4. FIES_ABOVE_HEATER_RADIATOR
5. FIES_ATMOSPHERIC_PRESSURE

“” Thank you for the update!

(...) I can say that we expect that the temperature near the grating, (...), would be a place where we expect large correlation with the drifts (RVs). FIES_GRATING_TANK_HOUSING is just by the grating, so this makes sense. The other temperatures might also make sense, such as FIES_INSIDE_BLACK_BOX_CENTRE and FIES_INSIDE_WHITE_BOX_CENTRE.

(...) surprised that FIES_ABOVE_HEATER_RADIATOR has such great impact, given that it is a very noisy instrument.

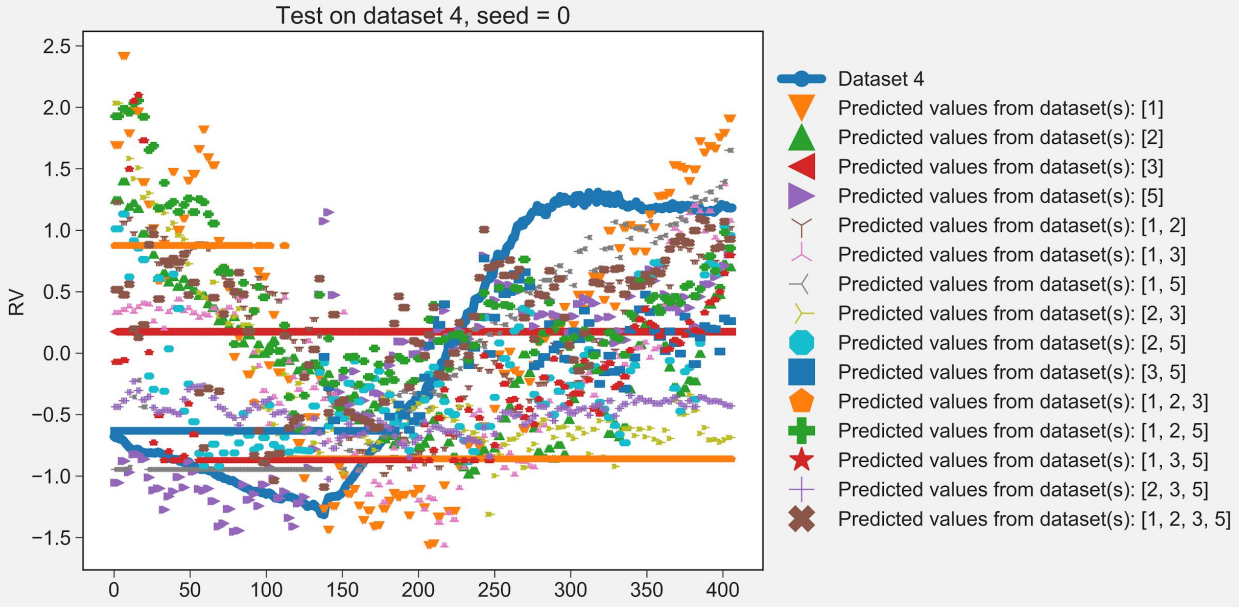
(...) On the other hand, it surprises me that FIES_ATMOSPHERIC_PRESSURE has any influence (...) in a pressure chamber (...) as we have a very expensive barometer!“”

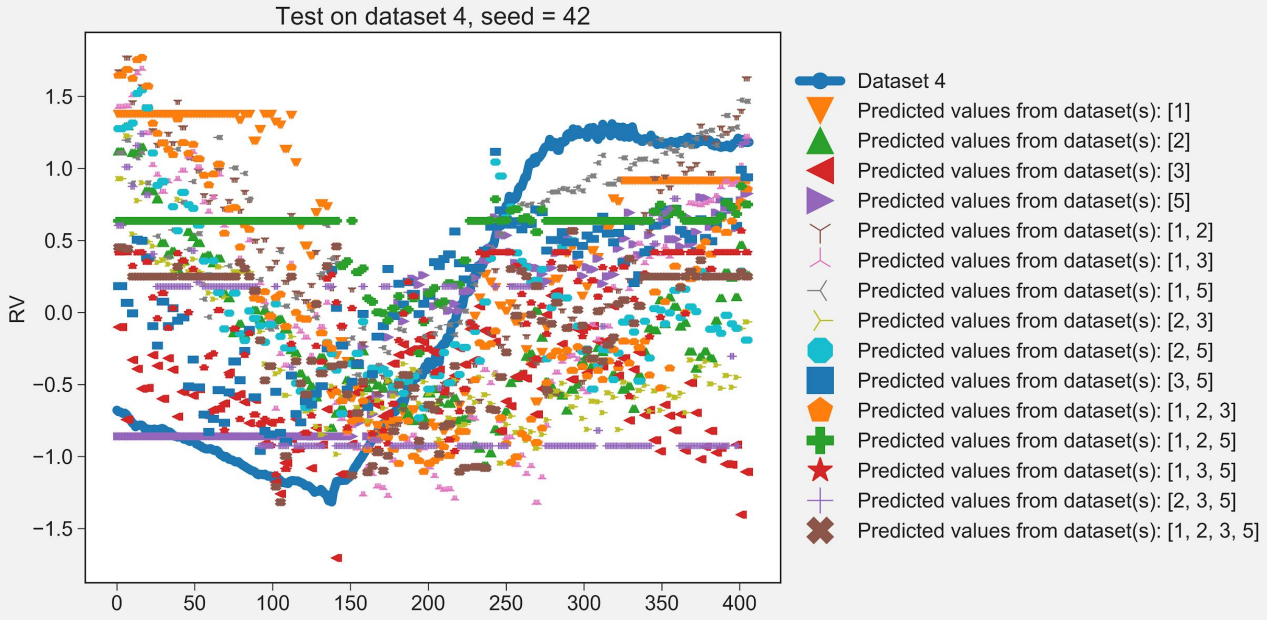
-Prof. Lars A. Buchhave, DTU Space

Telemetry data correlations

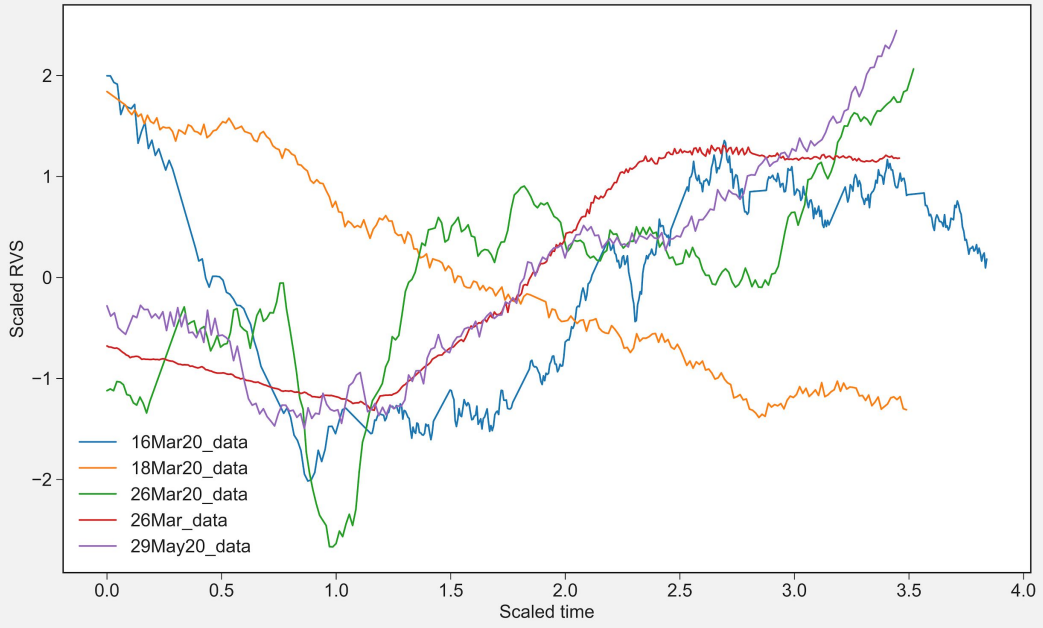
Importance of data sample size

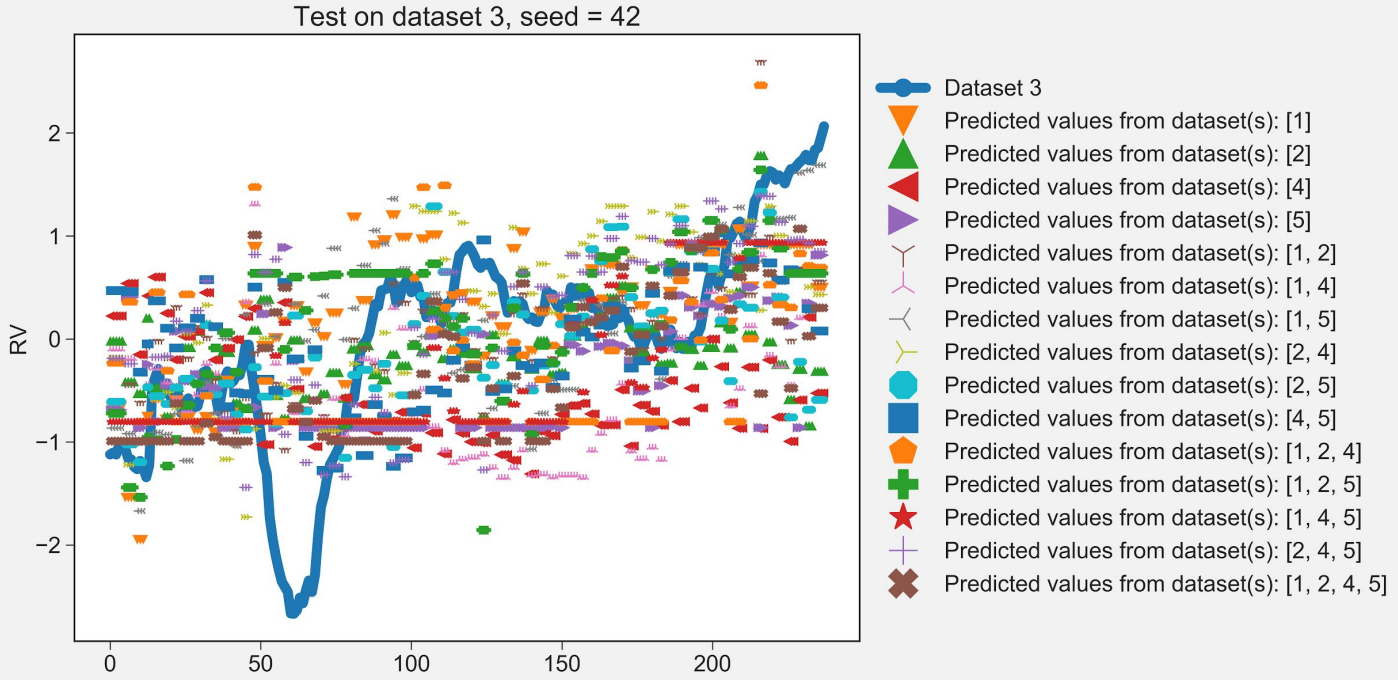
MLPR NN optimized with random search





Representative Data





MLPR NN optimized with random search

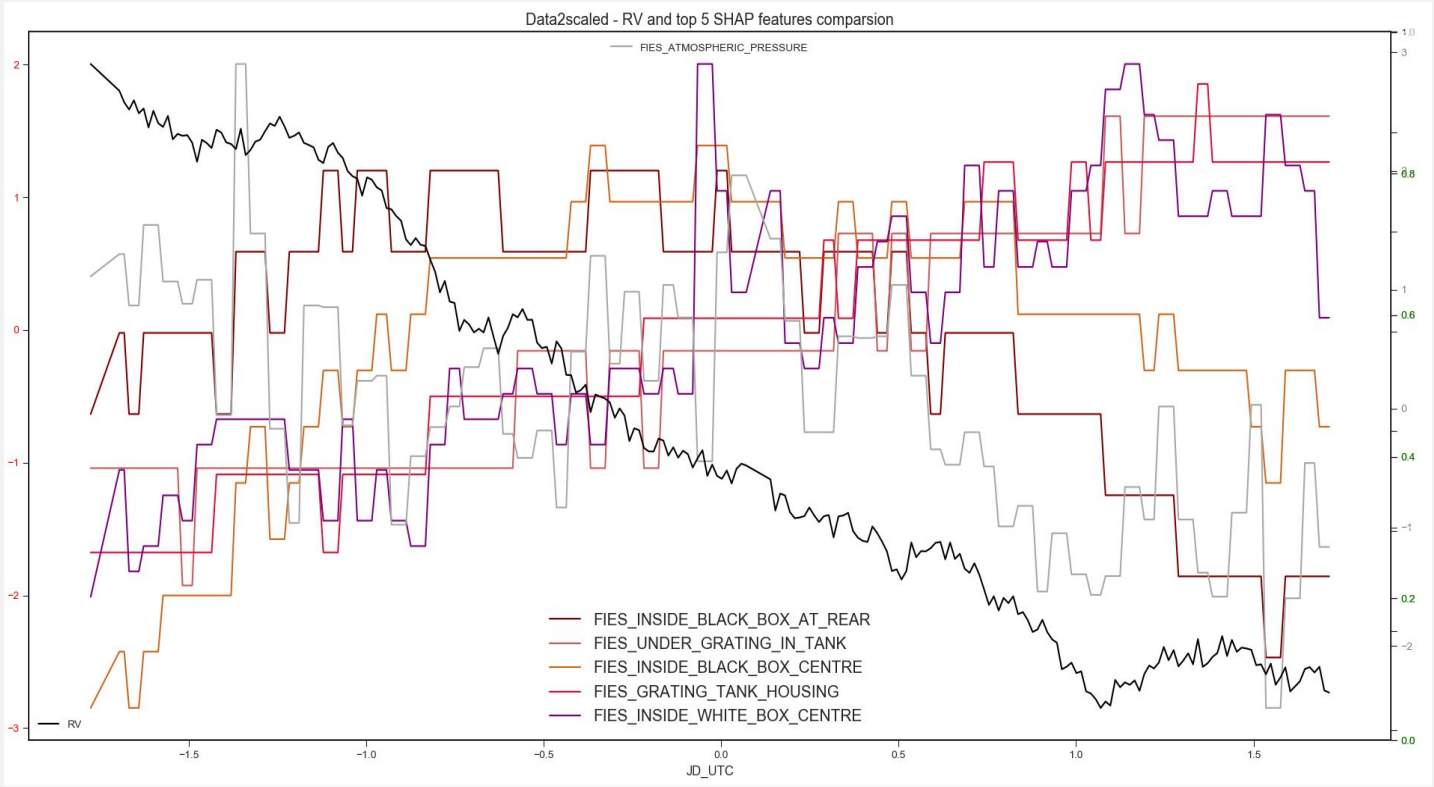
Conclusion

Data sample size makes predictions difficult

Confidence of ranked variables

Expectations compared to results

Further work



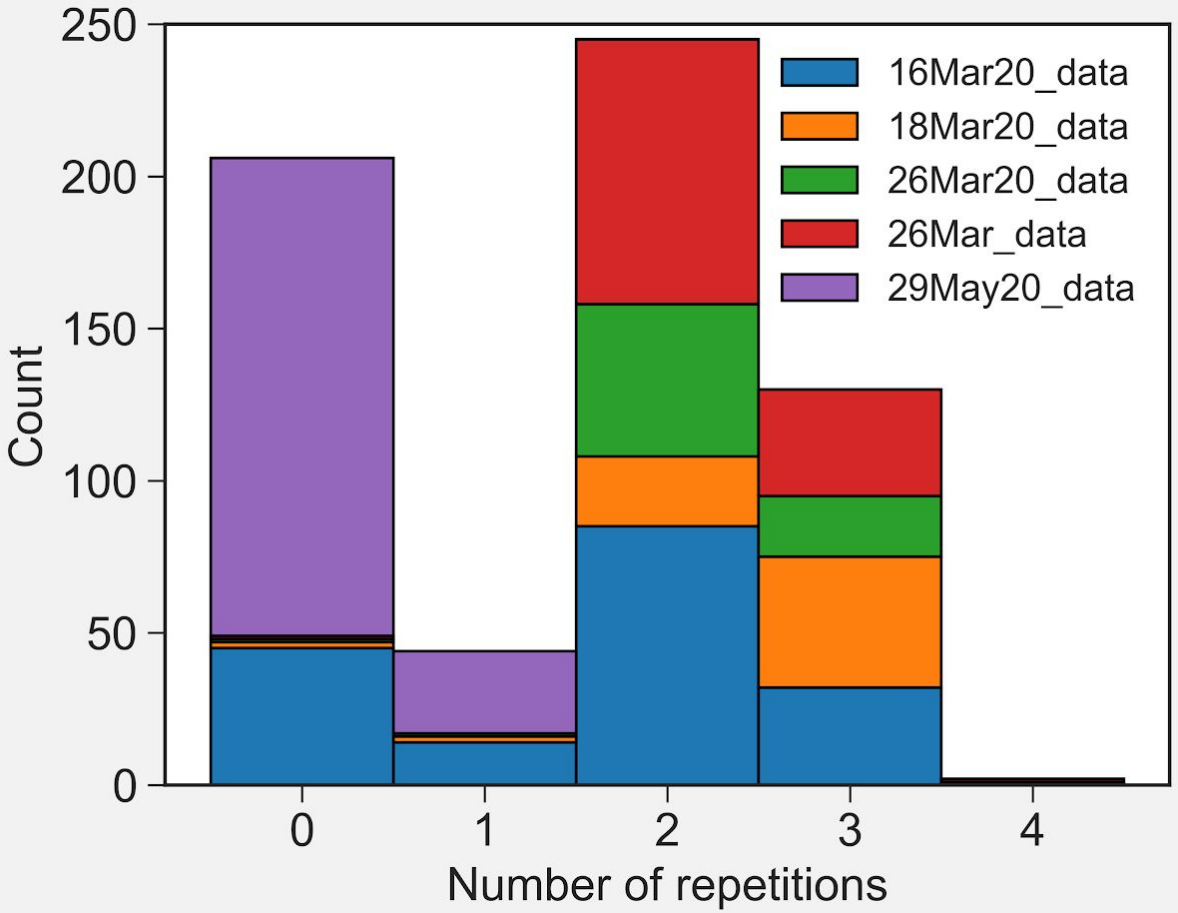
Appendix

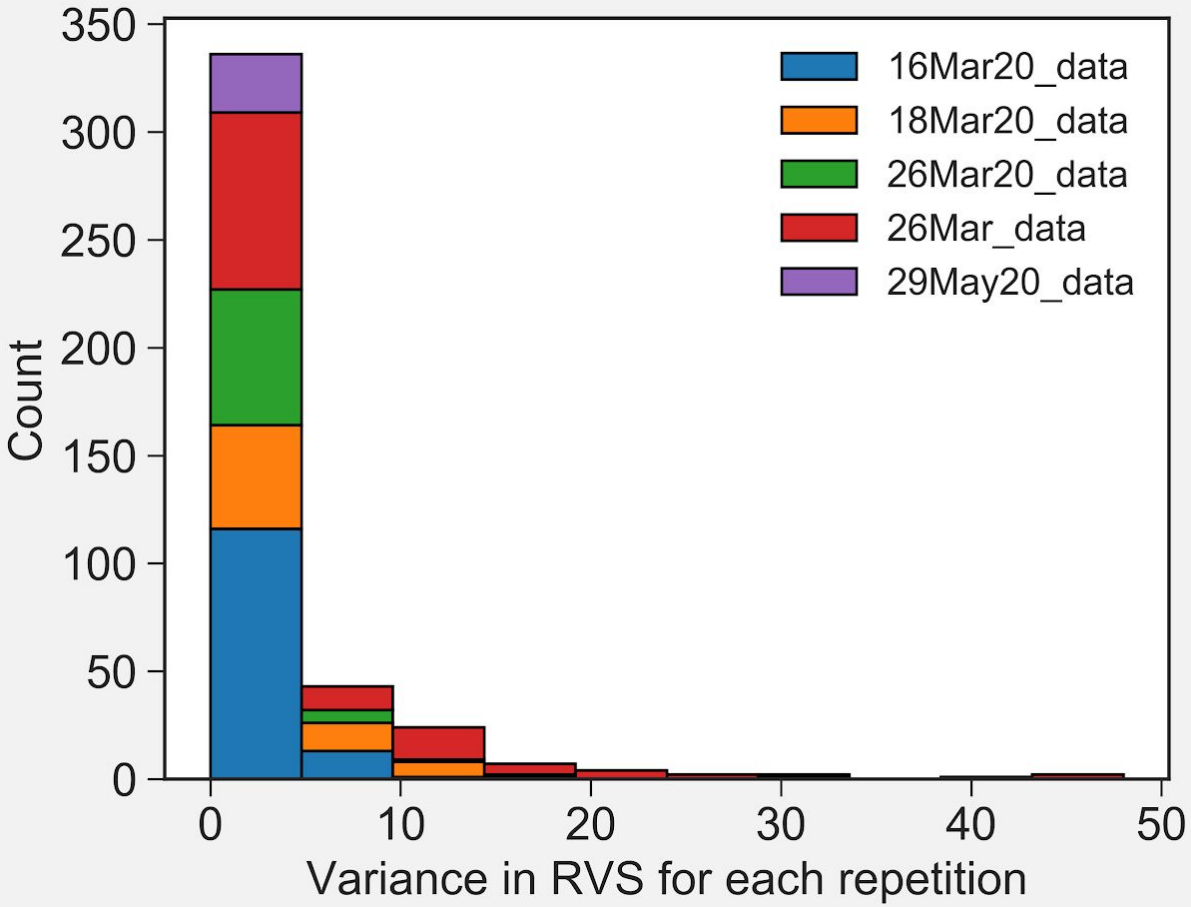
The scripts behind the scenes

To see all the code we wrote in order to get the presented results (and all the attempts that didn't produce any results), please refer to
our shared GitHub repo

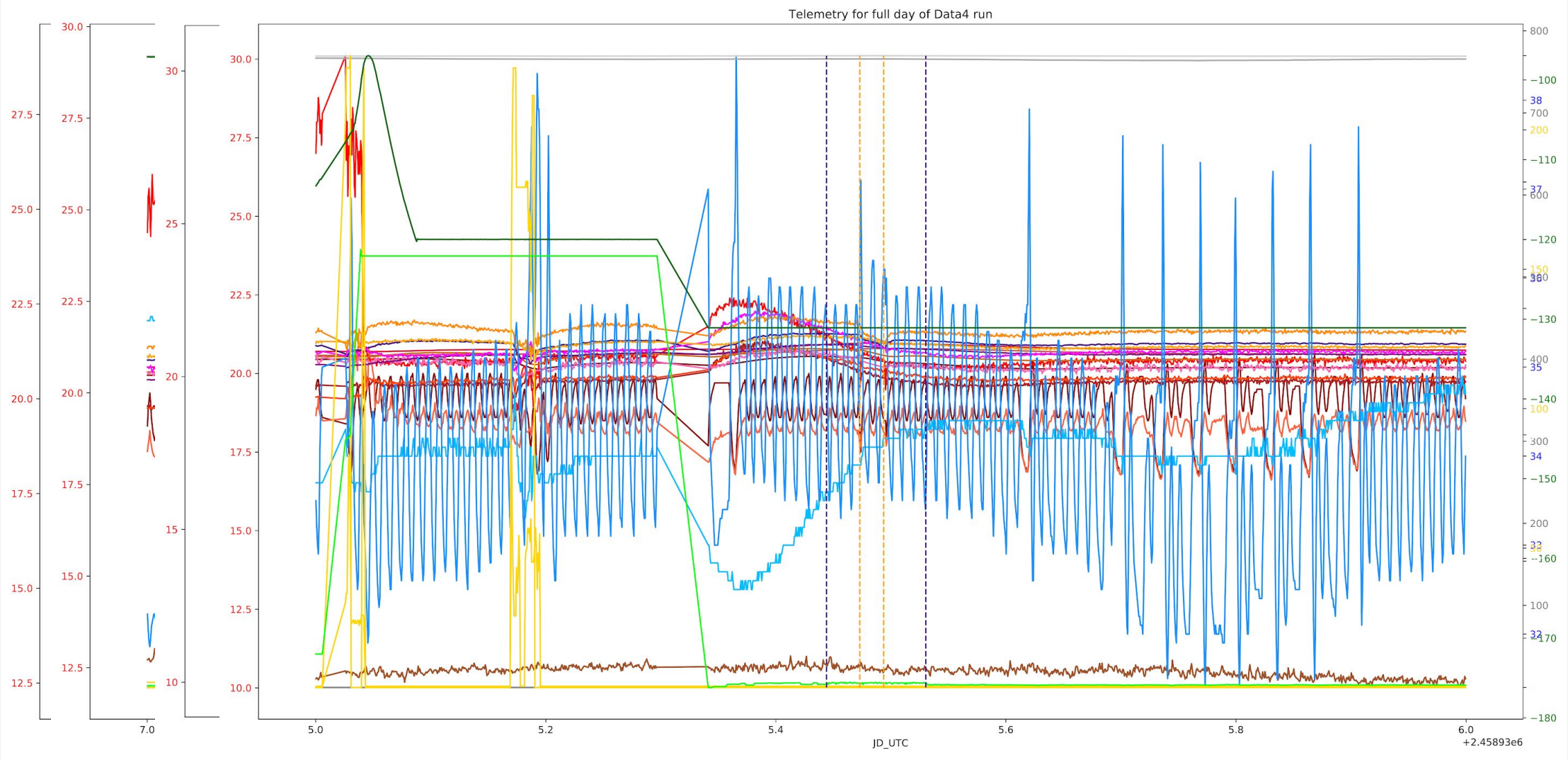
To run the LSTM efficiently, we have parallelized the scripts including this algorithm to a Google Colab GPU. An example is found here

Number of telemetry copies (repetitions)

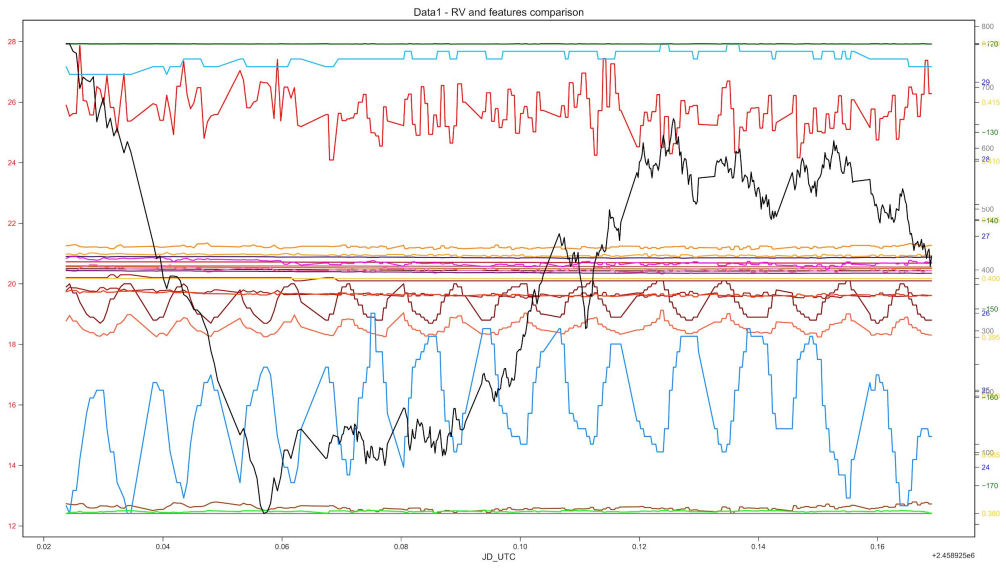


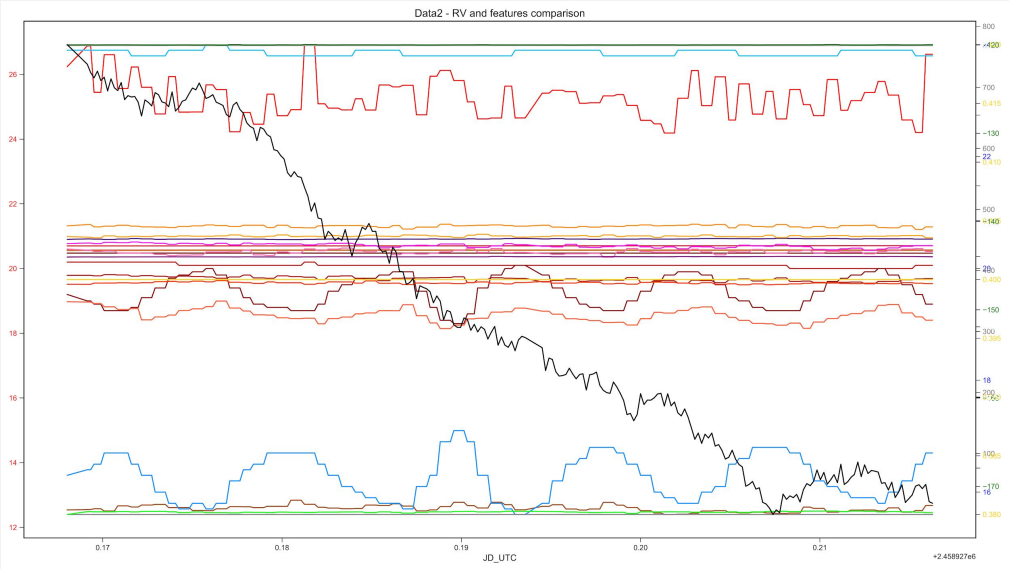


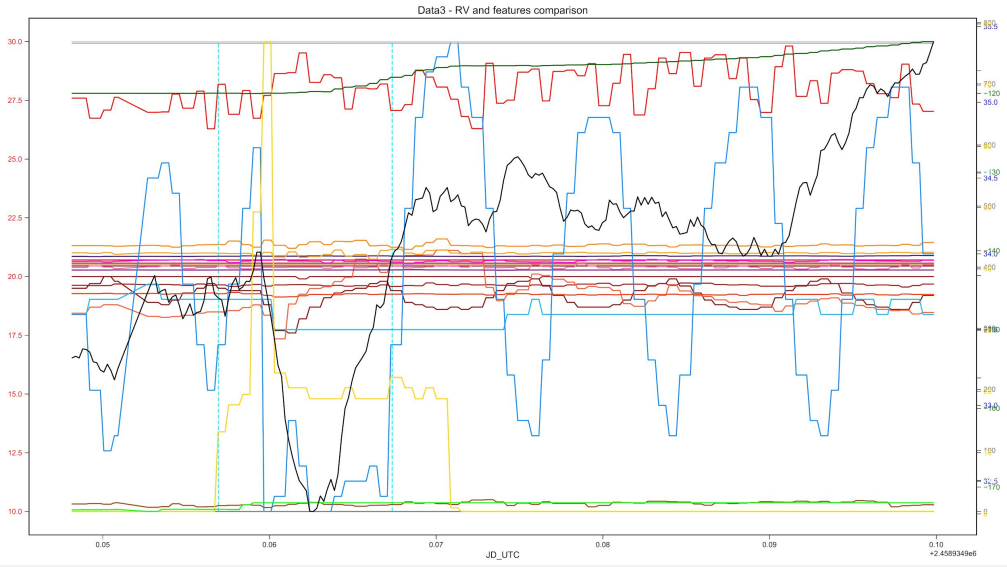
Visualization of the telemetry data

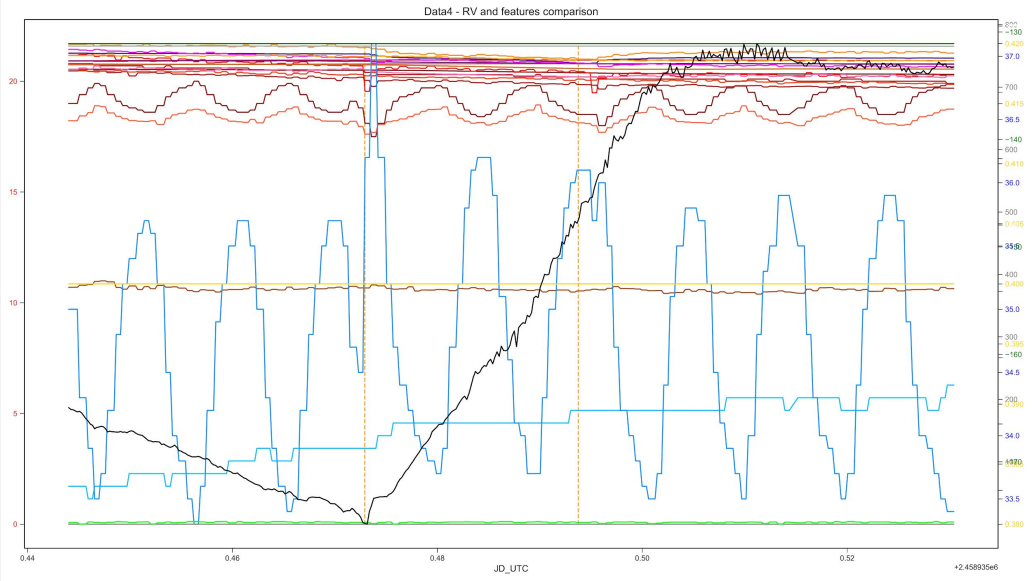


Visualization of the telemetry with RV data

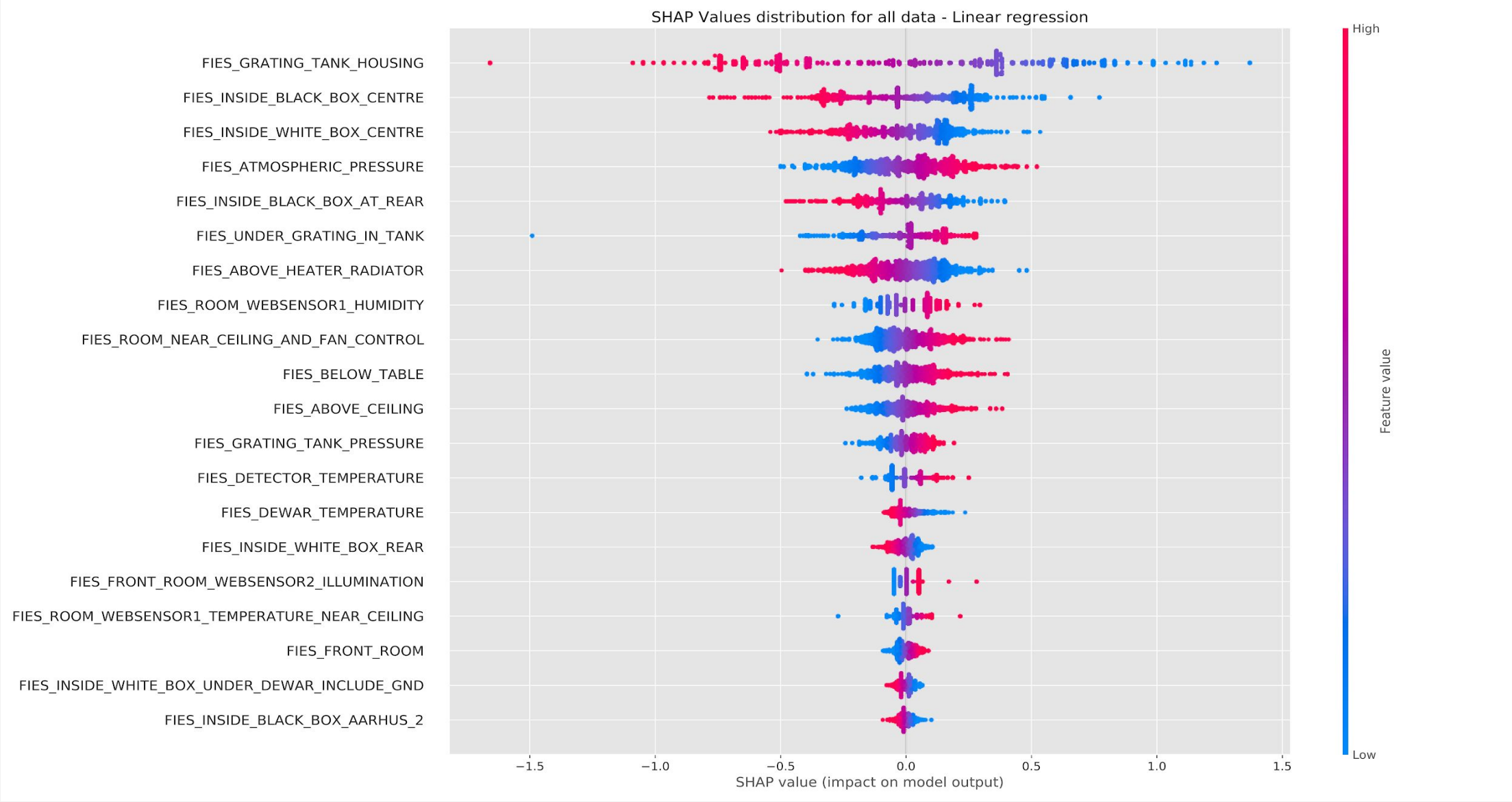




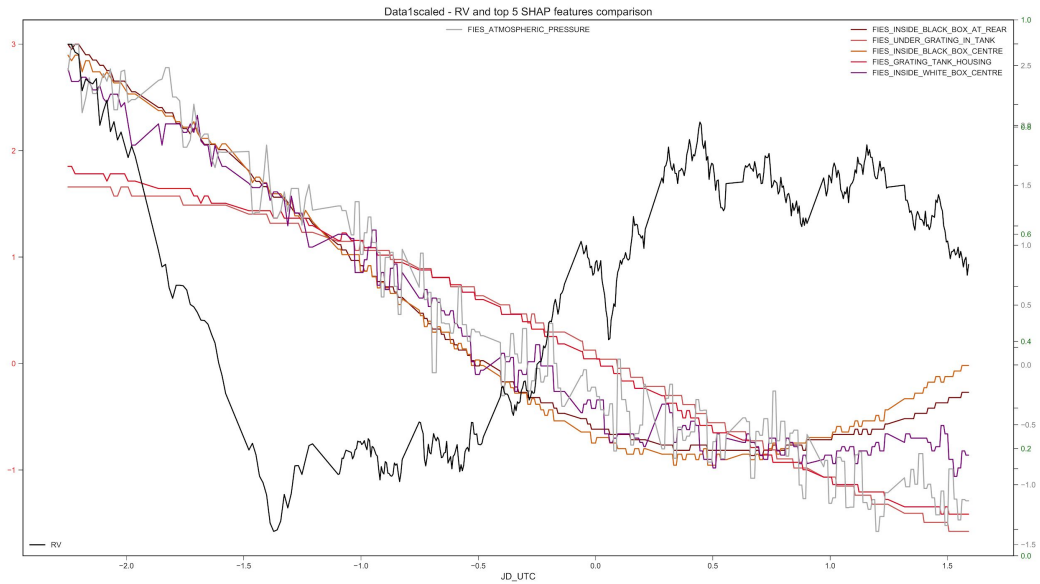


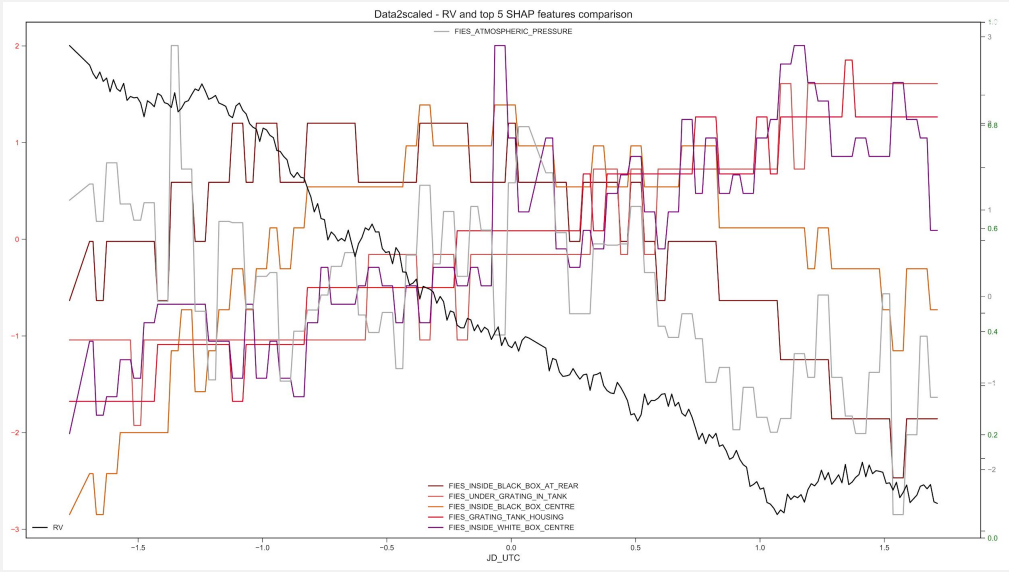


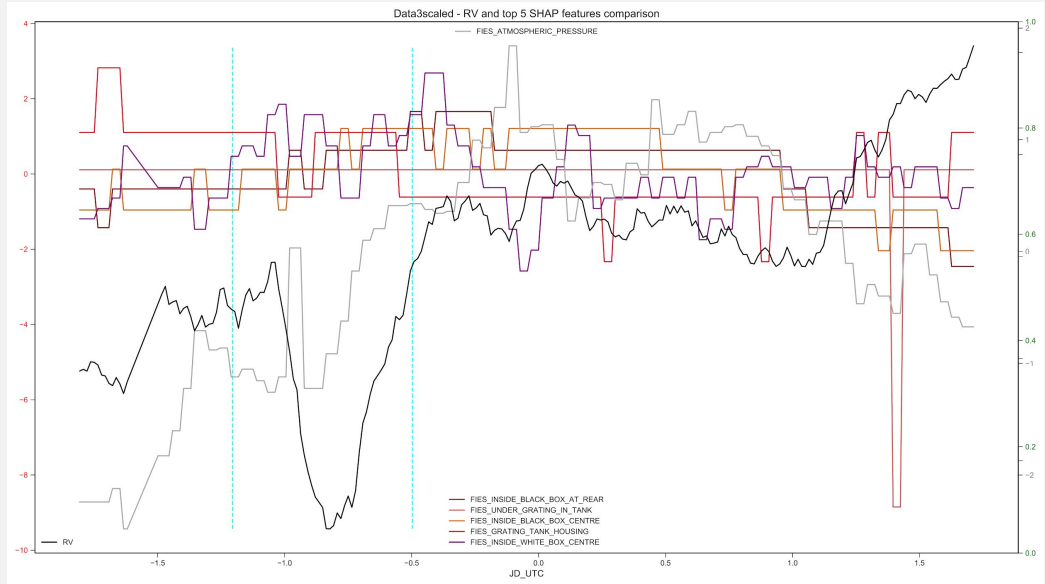
SHAP-values

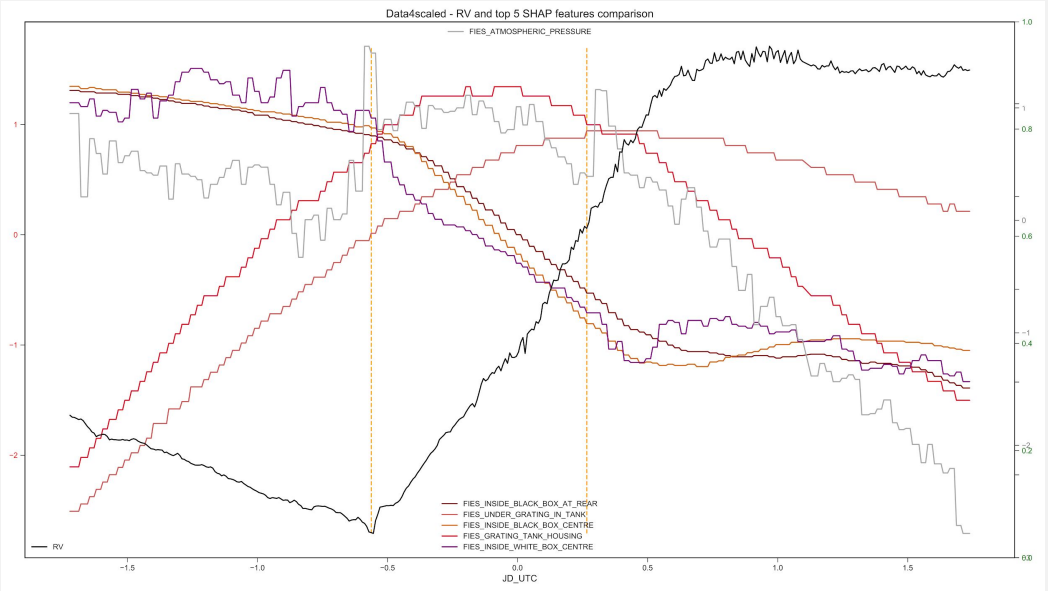

SHAP Values distribution for all data - Linear regression
FIES_GRATING_TANK_HOUSING
FIES_INSIDE_BLACK_BOX_CENTRE
FIES_INSIDE_WHITE_BOX_CENTRE
FIES_ATMOSPHERIC_PRESSURE
FIES_INSIDE_BLACK_BOX_AT_REAR
FIES_UNDER_GRATING_IN_TANK
FIES_ABOVE_HEATER_RADIATOR
FIES_ROOM_WEBSENSOR1_HUMIDITY
FIES_ROOM_NEAR_CEILING_AND_FAN_CONTROL
FIES_BELOW_TABLE
FIES_ABOVE_CEILING
FIES_GRATING_TANK_PRESSURE
FIES_DETECTOR_TEMPERATURE
FIES_DEWAR_TEMPERATURE
FIES_INSIDE_WHITE_BOX_REAR
FIES_FRONT_ROOM_WEBSENSOR2_ILLUMINATION
FIES_ROOM_WEBSENSOR1_TEMPERATURE_NEAR_CEILING
FIES_FRONT_ROOM
FIES_INSIDE_WHITE_BOX_UNDER_DEWAR_INCLUDE_GND
FIES_INSIDE_BLACK_BOX_AARHUS_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
SHAP value (impact on model output)
High
Feature value
Low

Visualization of the scaled telemetry and RV - Top 5 SHAP features

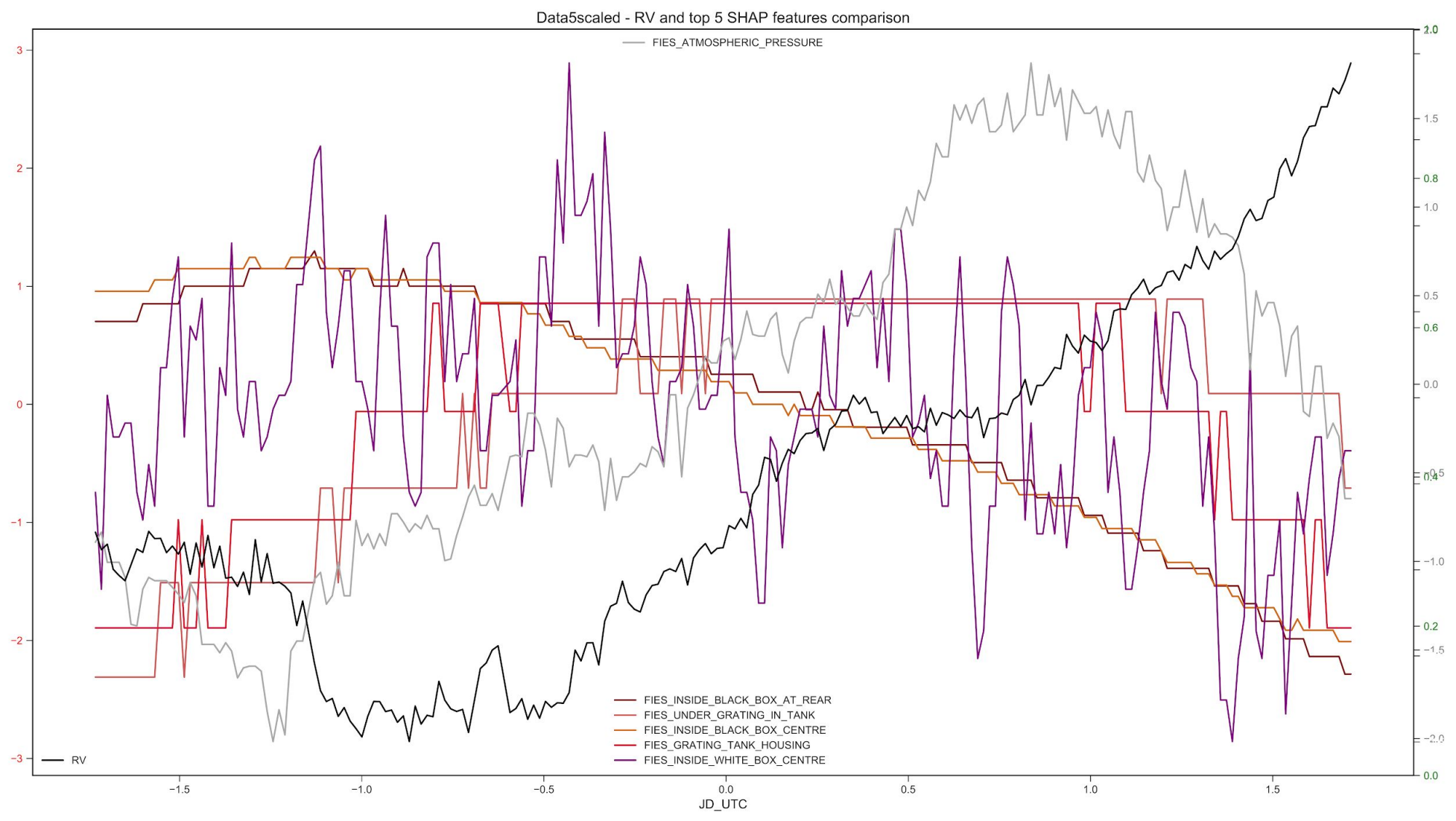




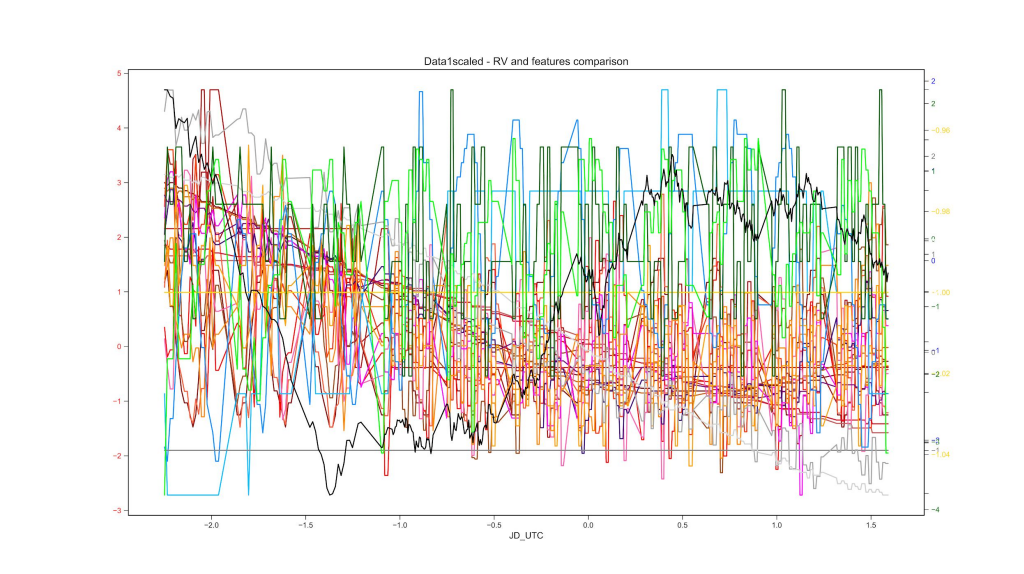


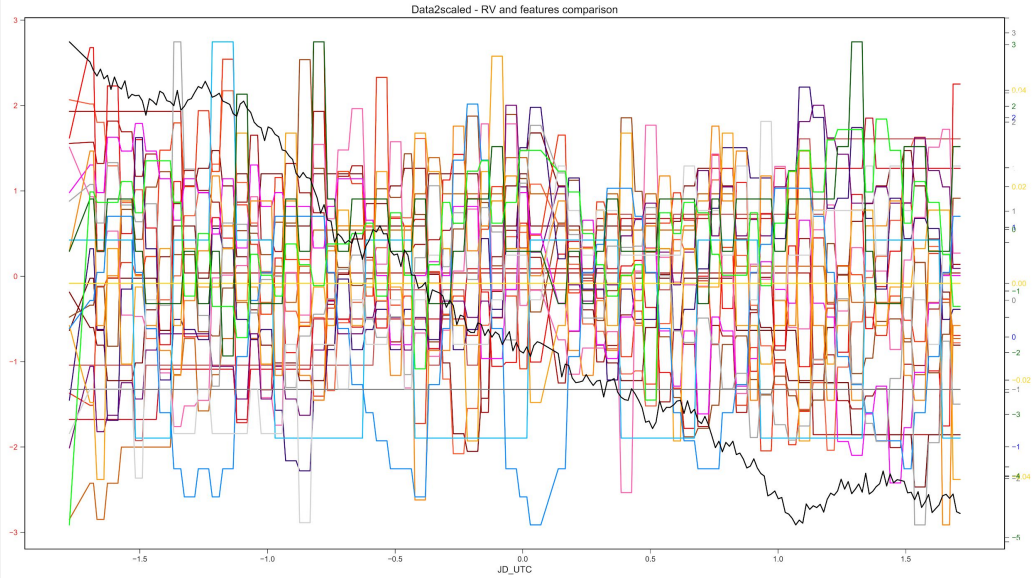


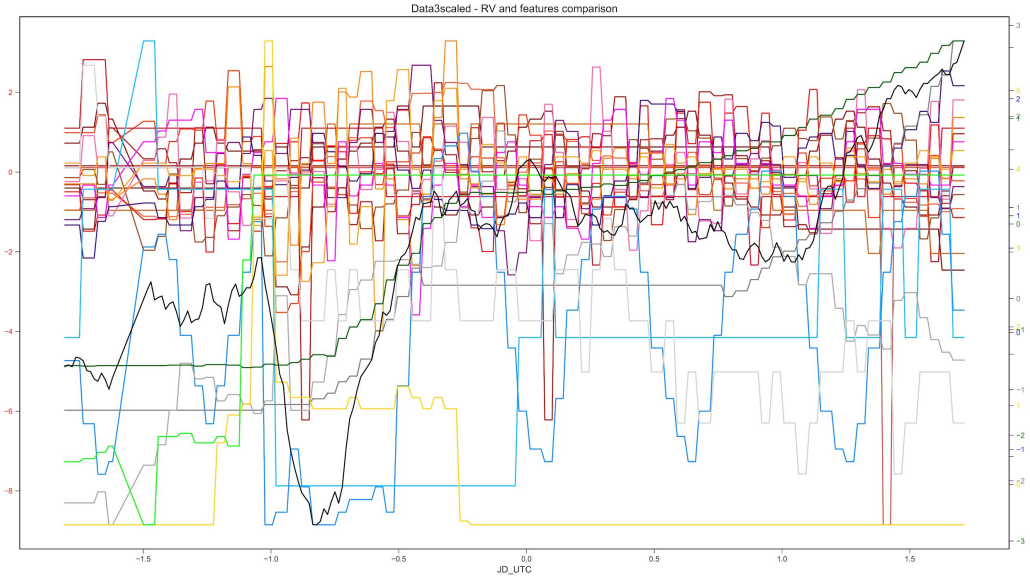
Visualization of the scaled telemetry and RV - Top 5 SHAP features

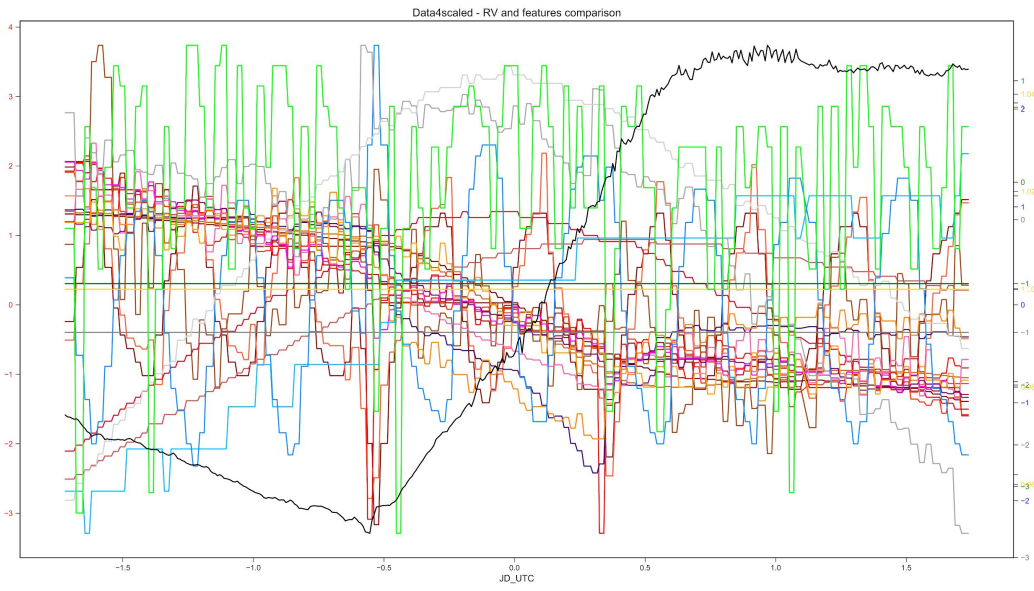


Visualization of the scaled telemetry with scaled RV data

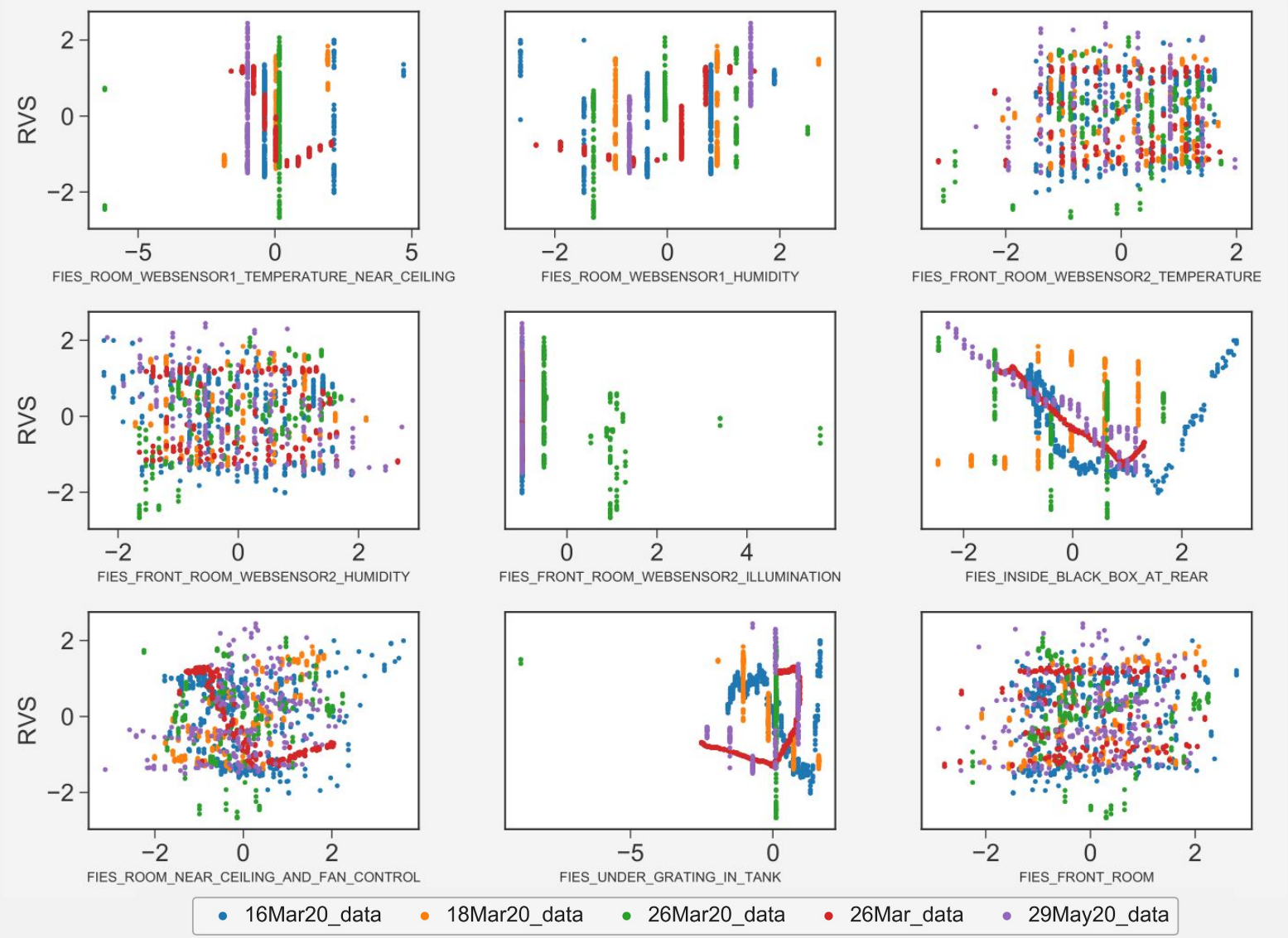




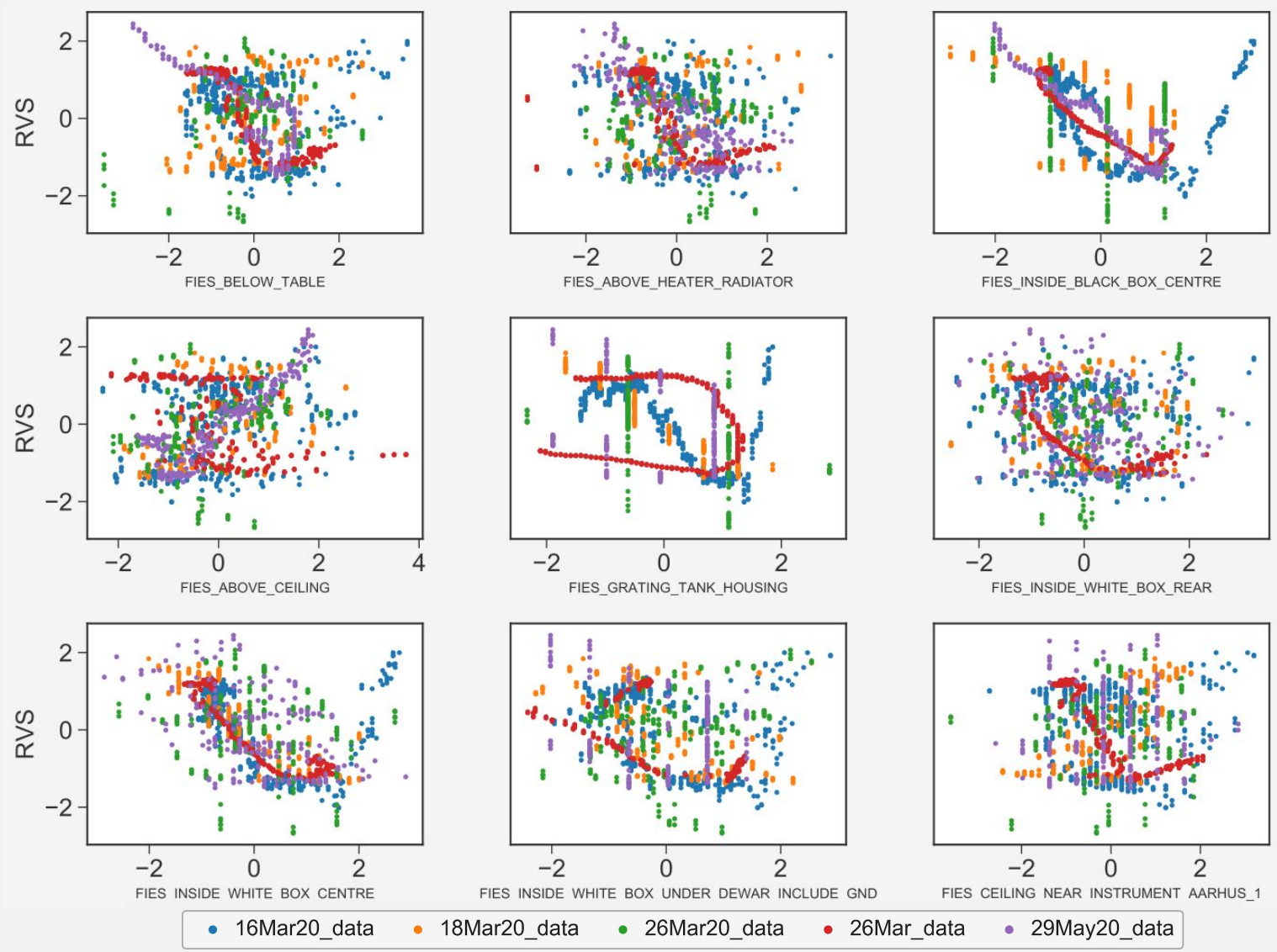




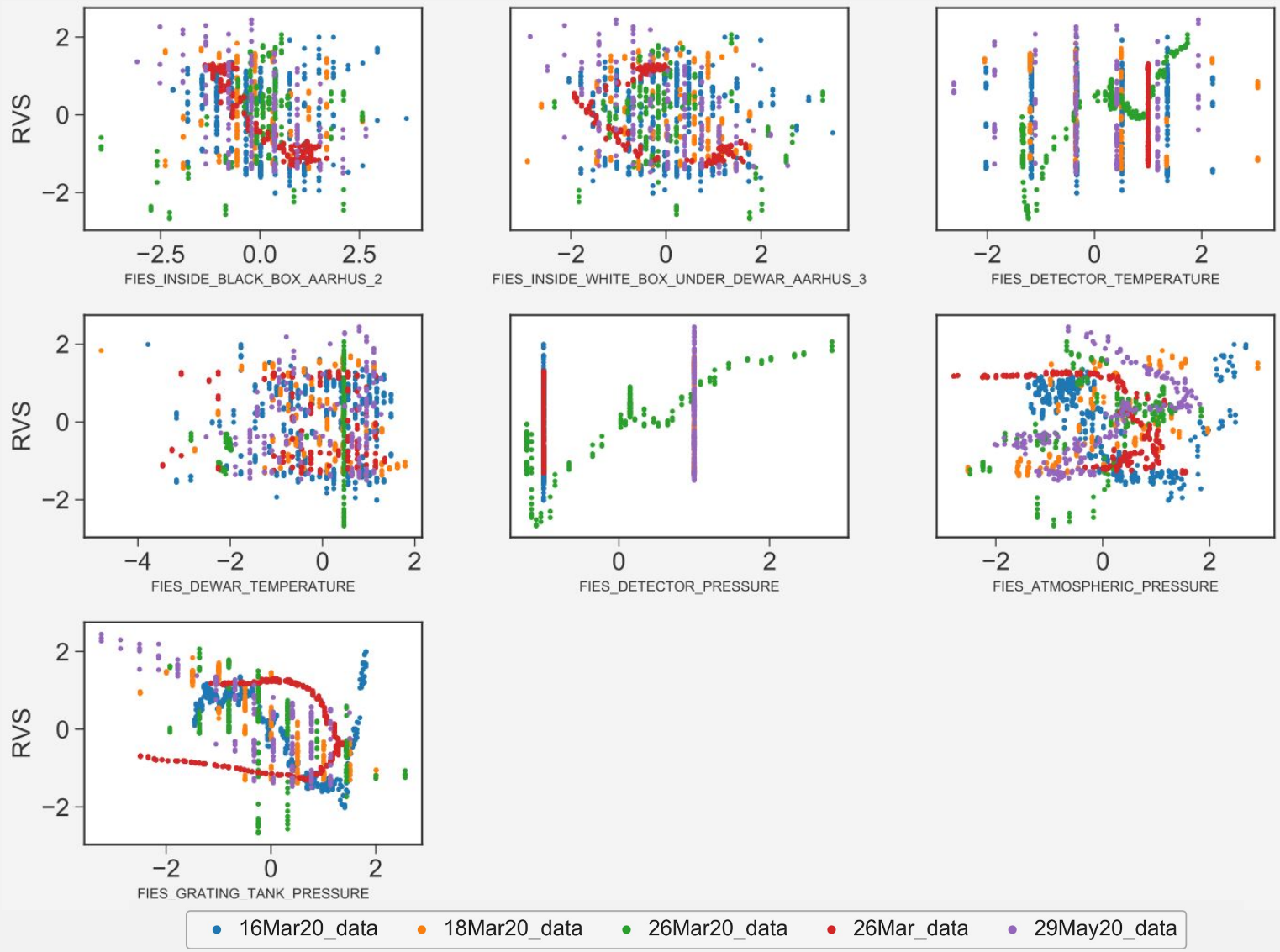
Correlations between RVS and parameters (Scaled) - Part 1/3



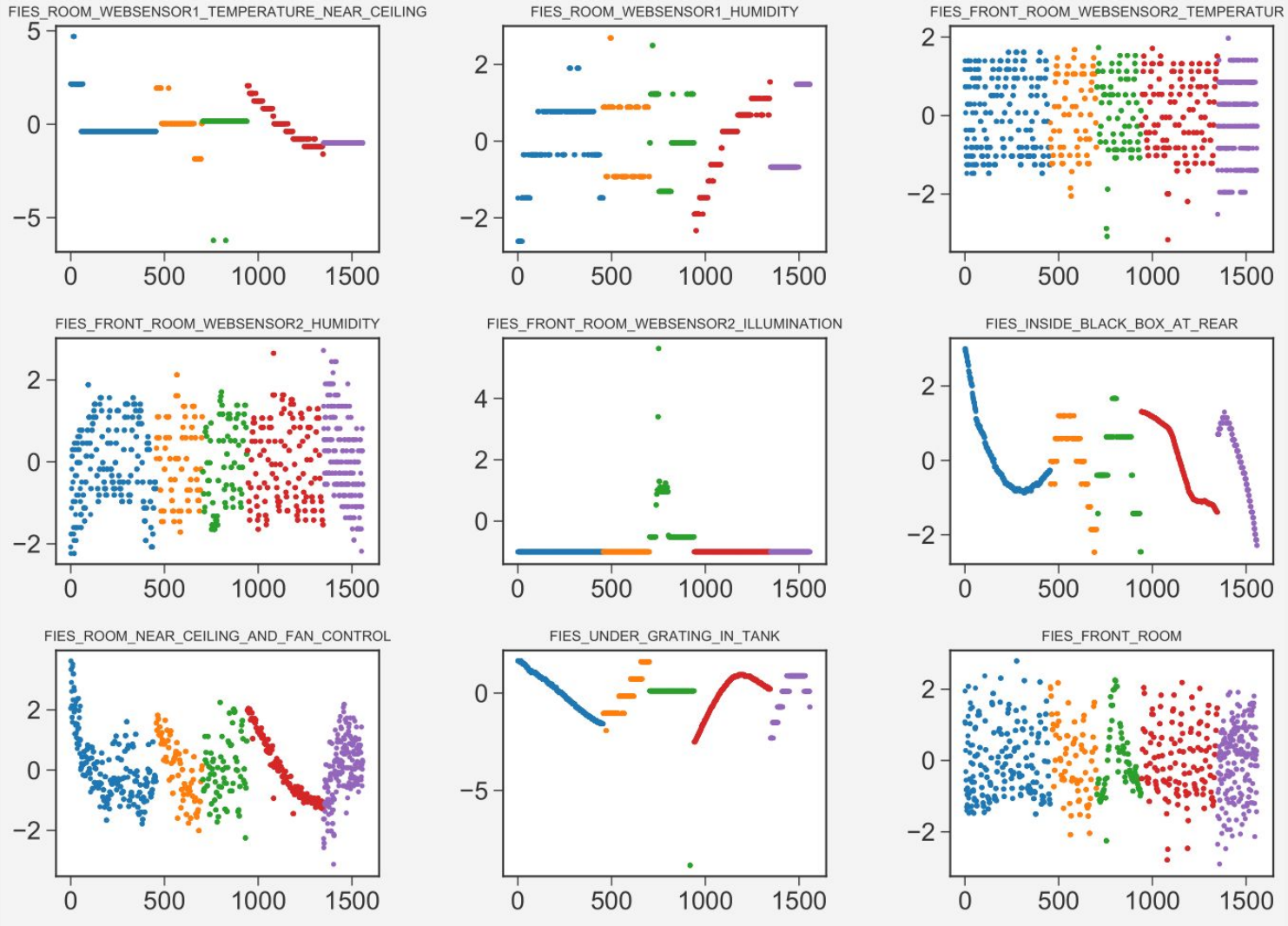
Correlations between RVS and parameters (Scaled) - Part 2/3



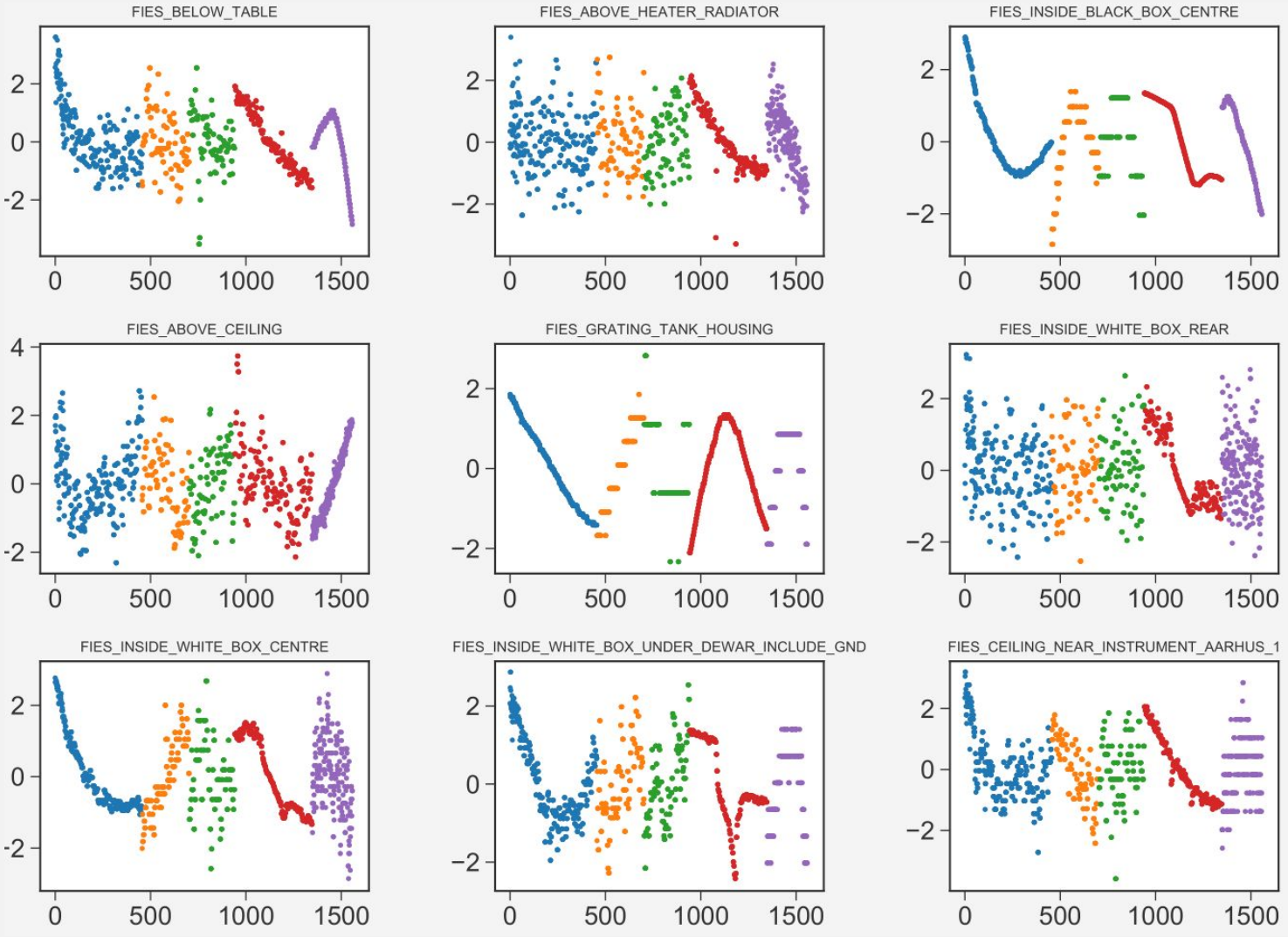
Correlations between RVS and parameters (Scaled) - Part 3/3



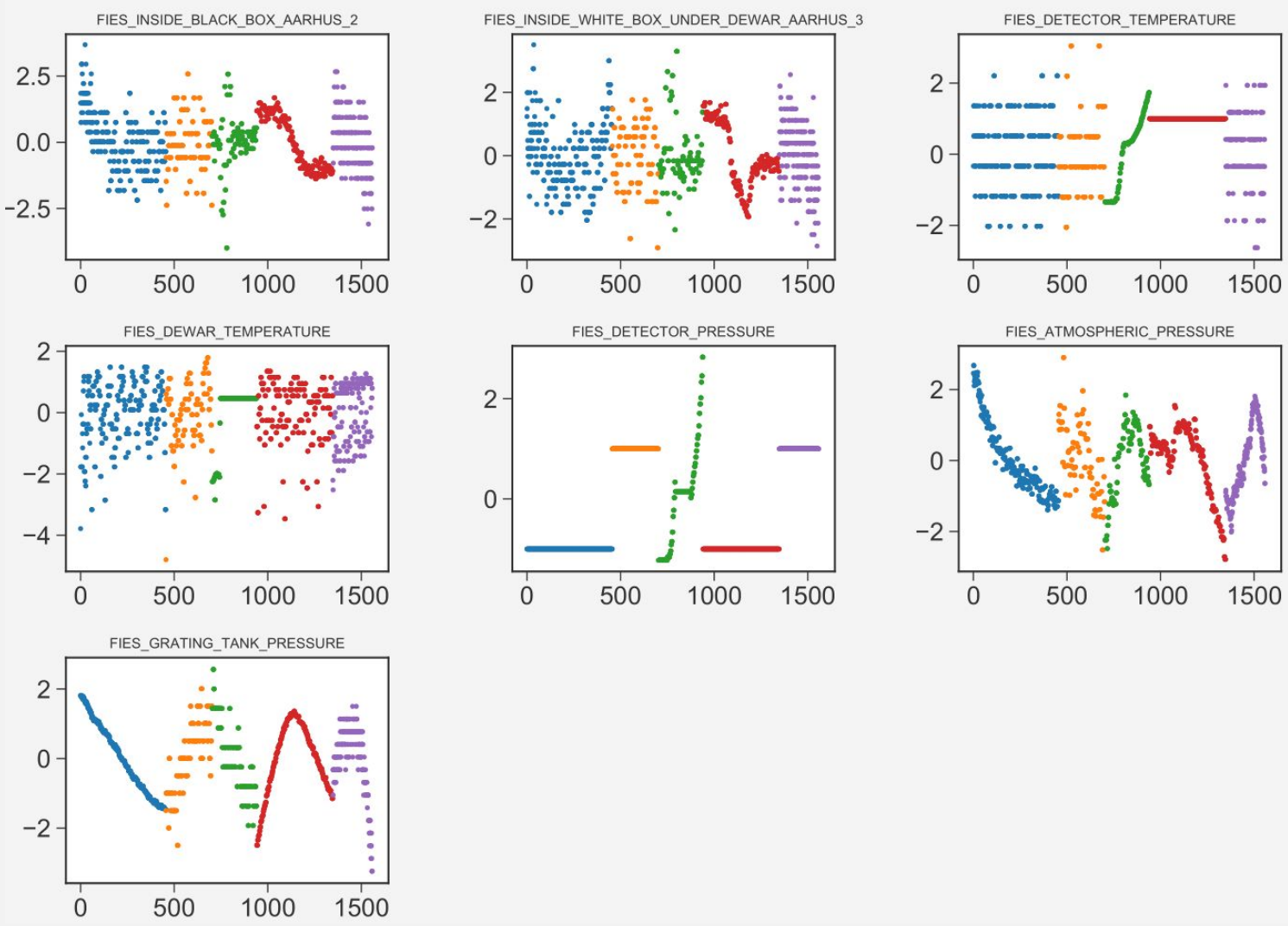
RVS vs. Time (Scaled) - Part 1/3



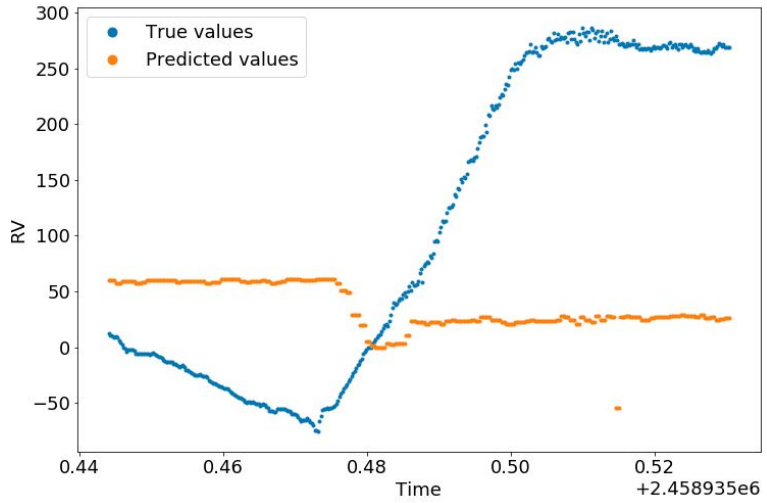
RVS vs. Time (Scaled) - Part 2/3

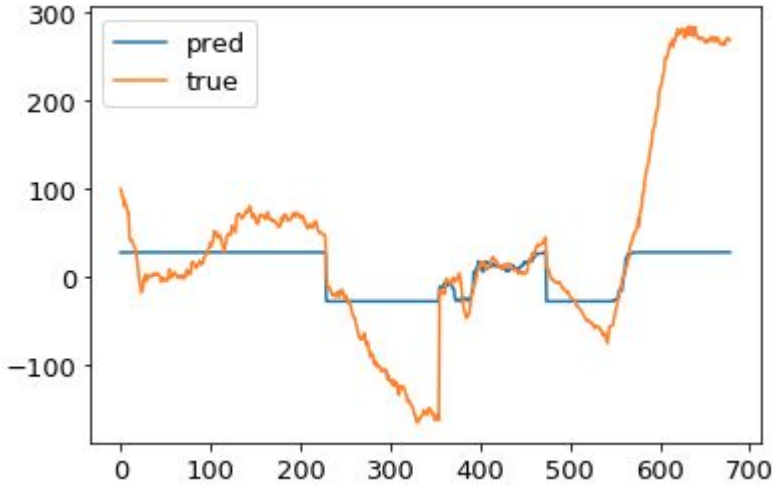


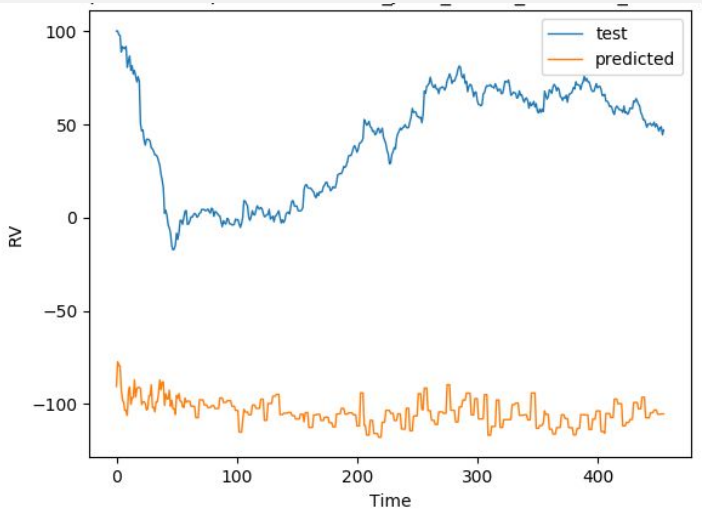
RVS vs. Time (Scaled) - Part 3/3

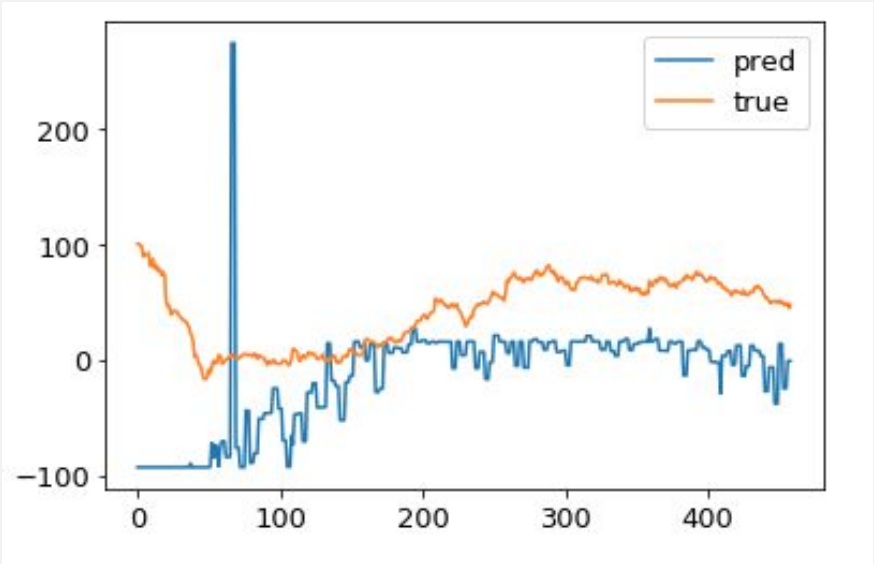


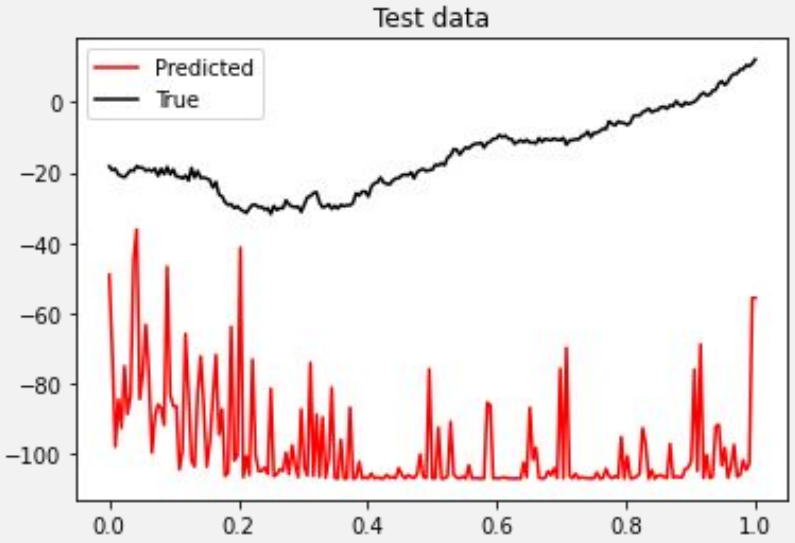
Different early failed attempts at machine learning



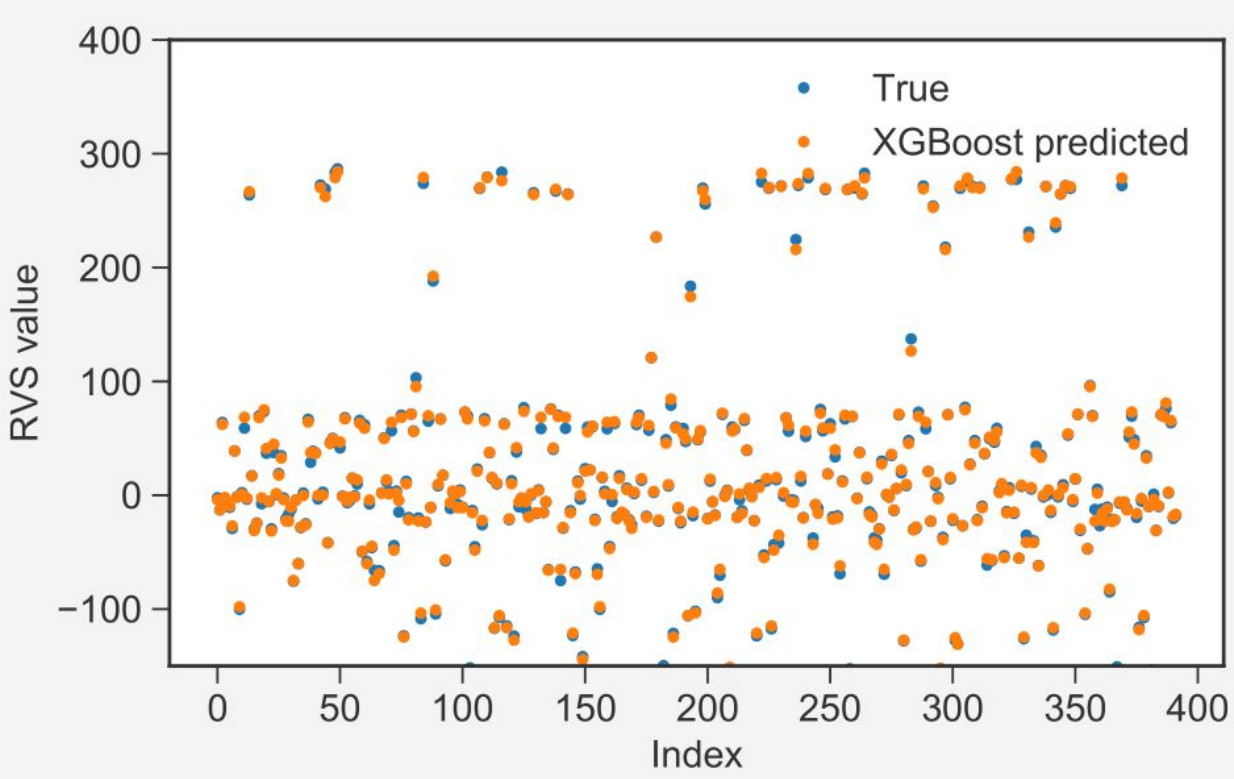


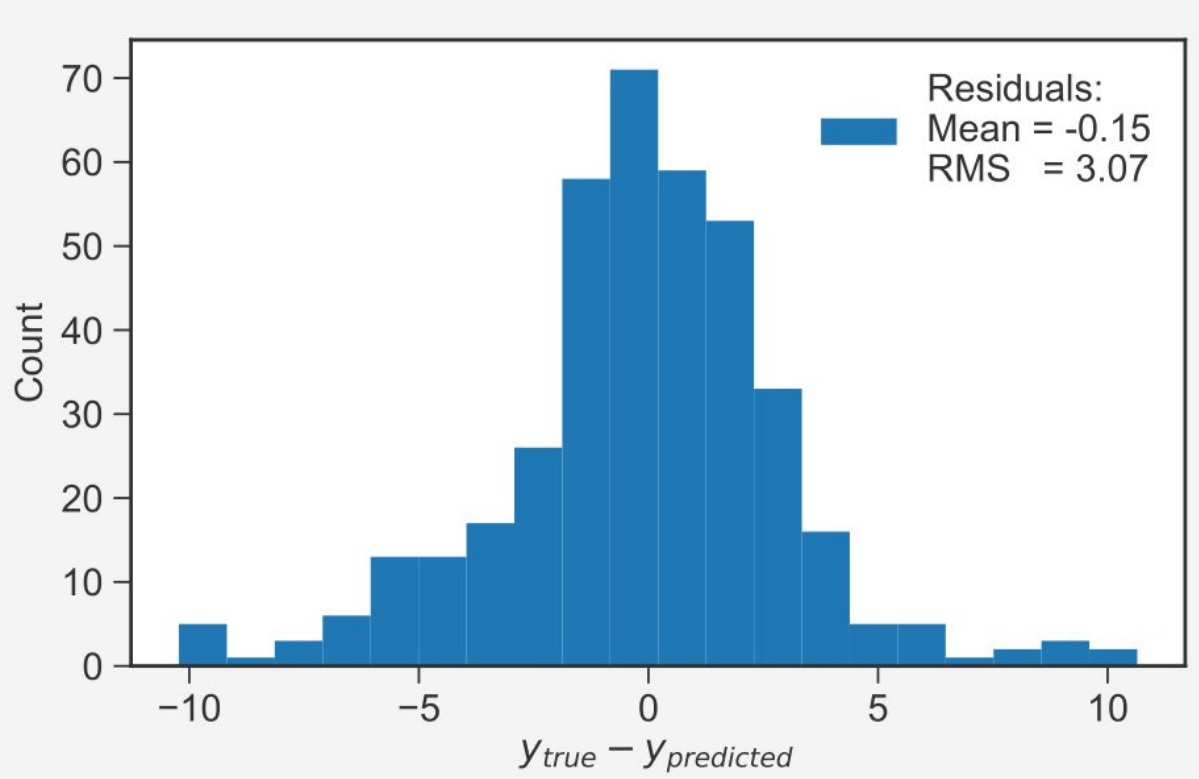




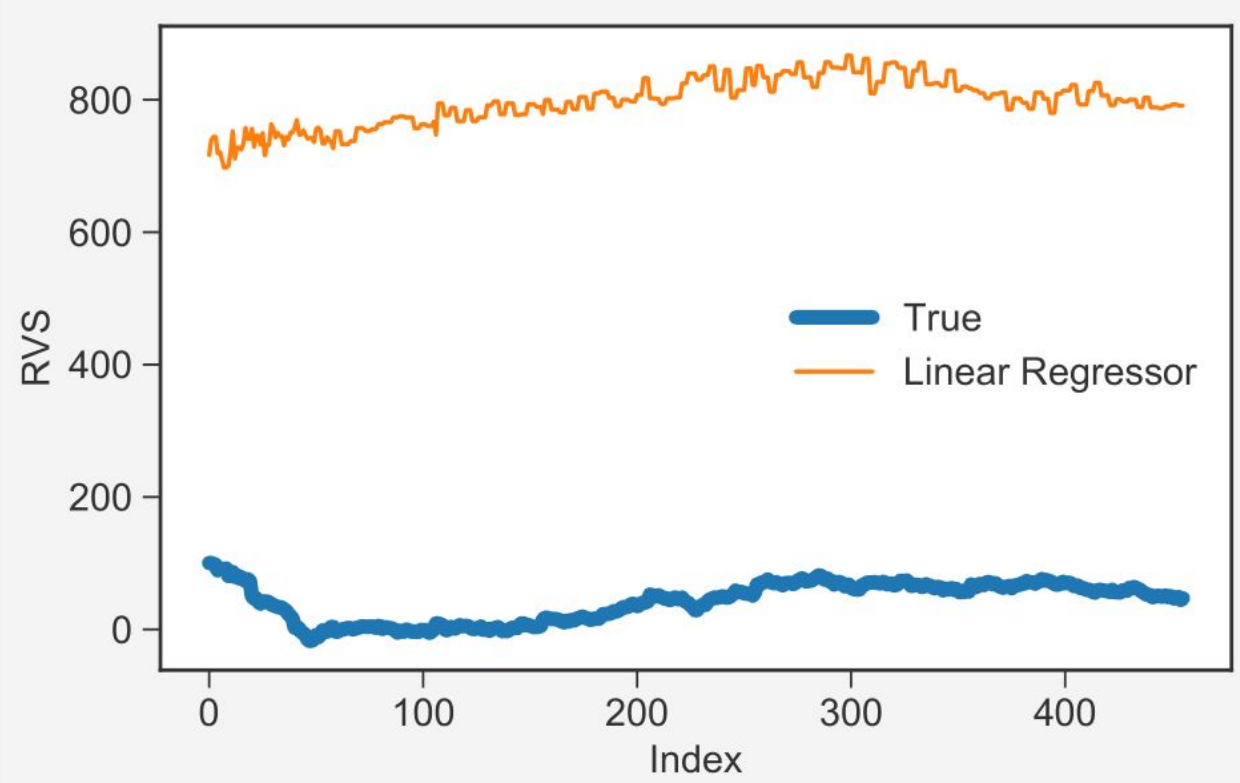


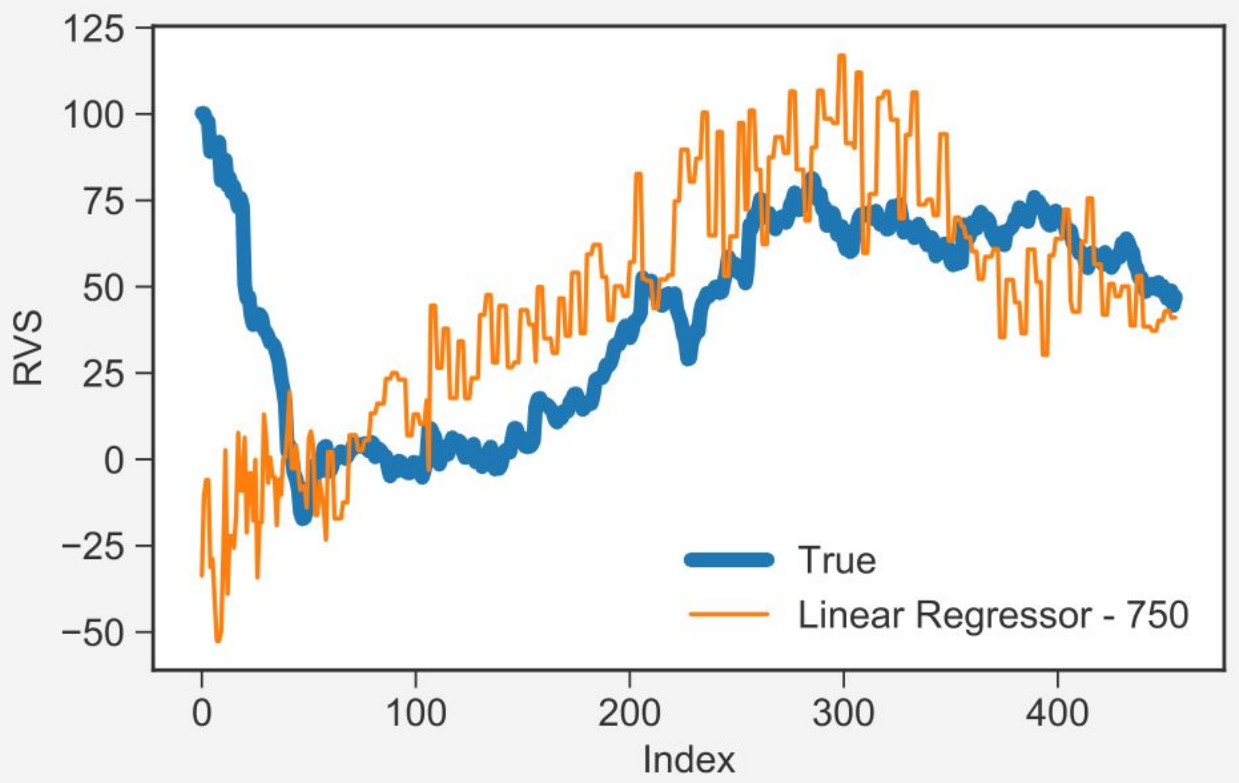
Predictions from randomly sampling train and test data from all sets using XGBoost (non-scaled)



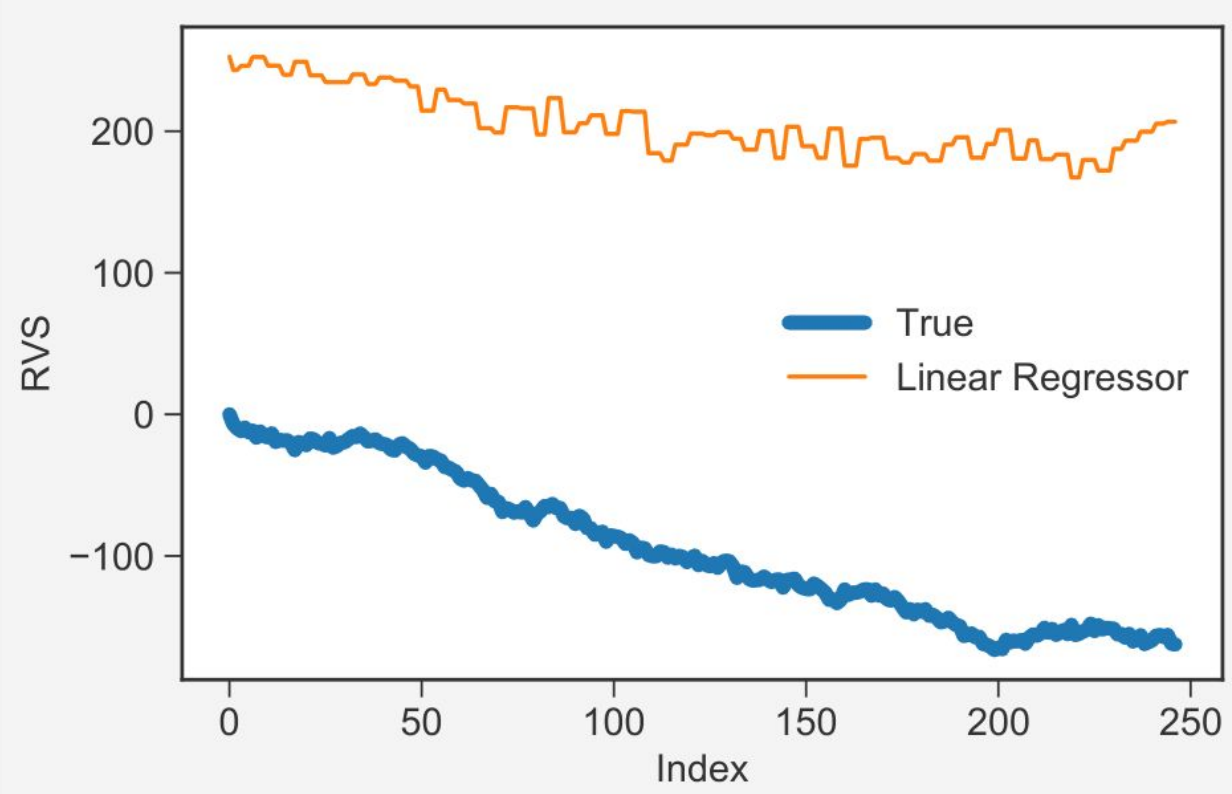


Predictions before scaling
LinReg on test 1/5

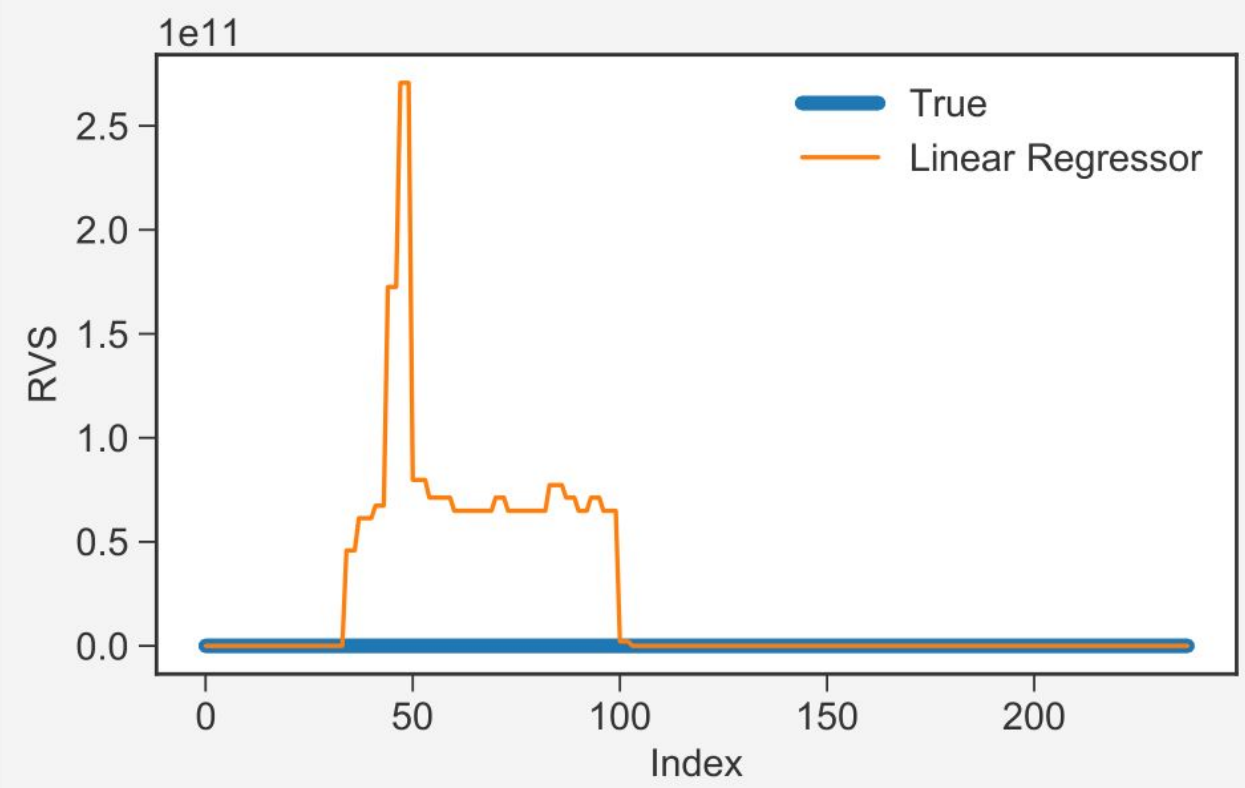


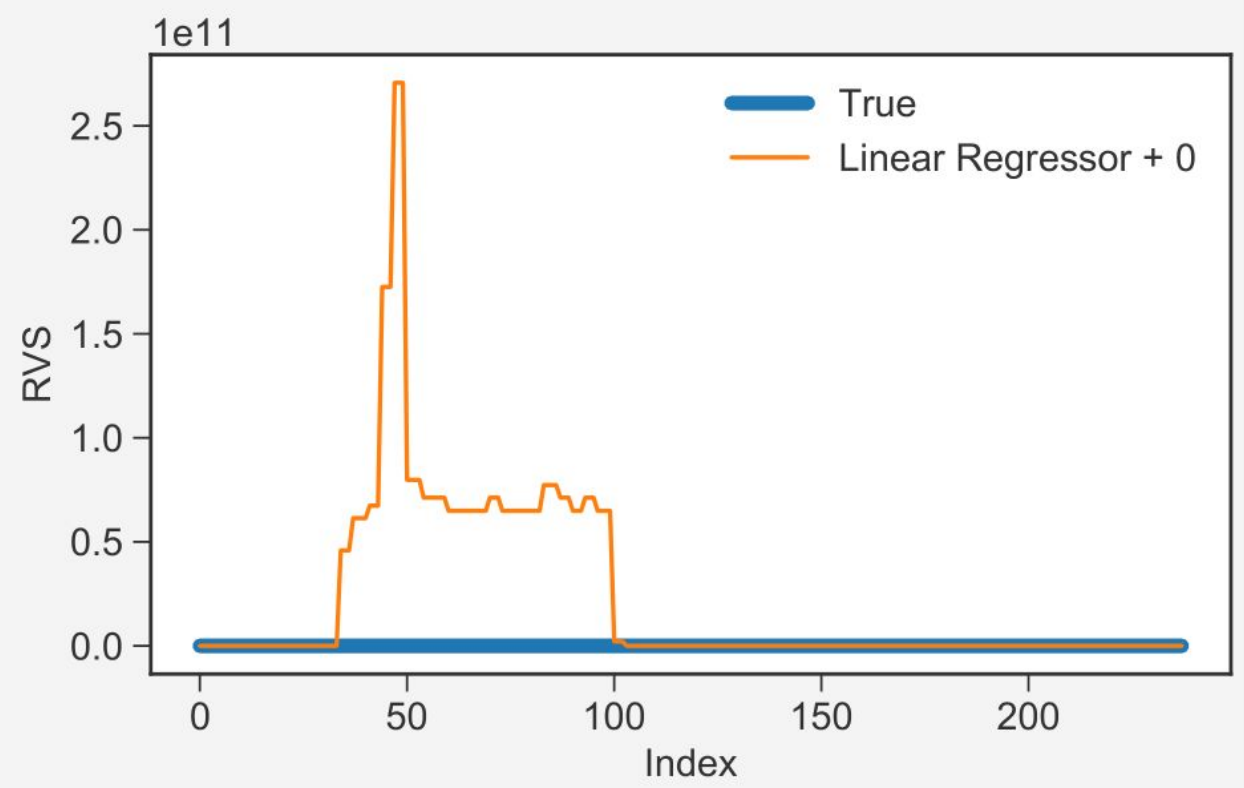


Predictions before scaling
LinReg on test 2/5

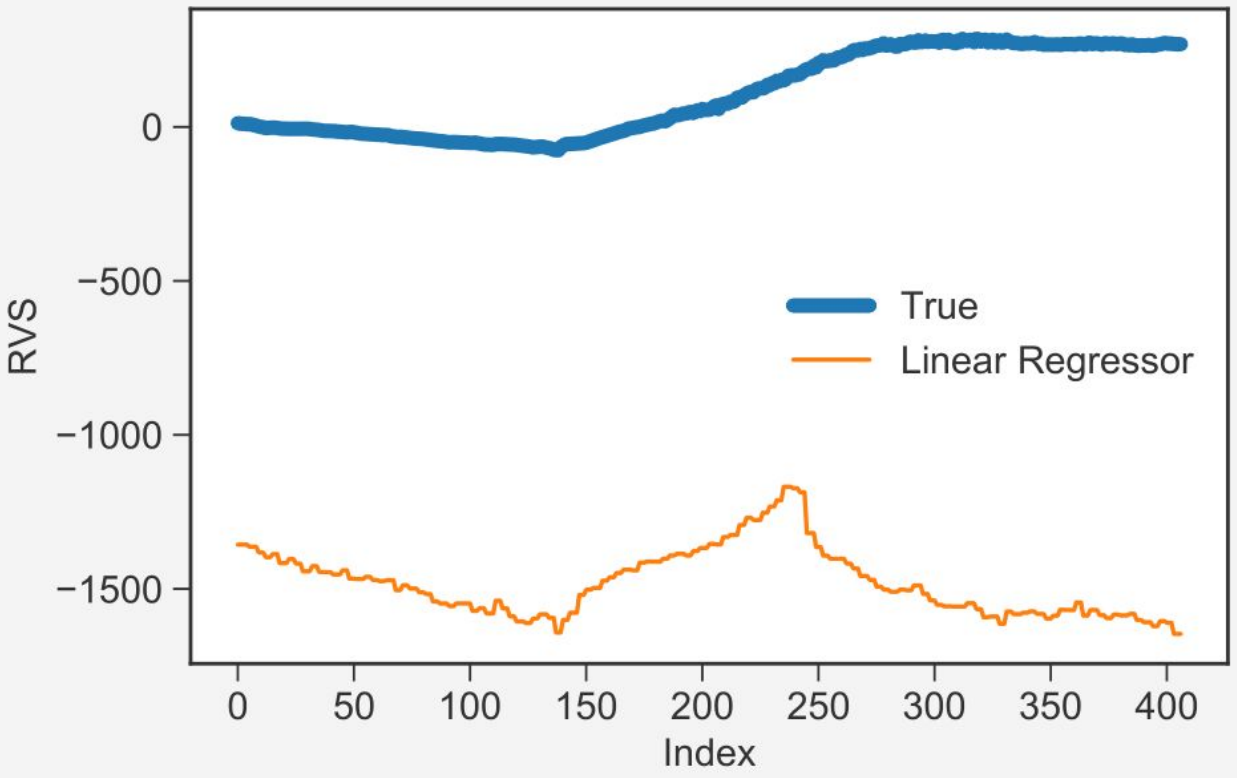


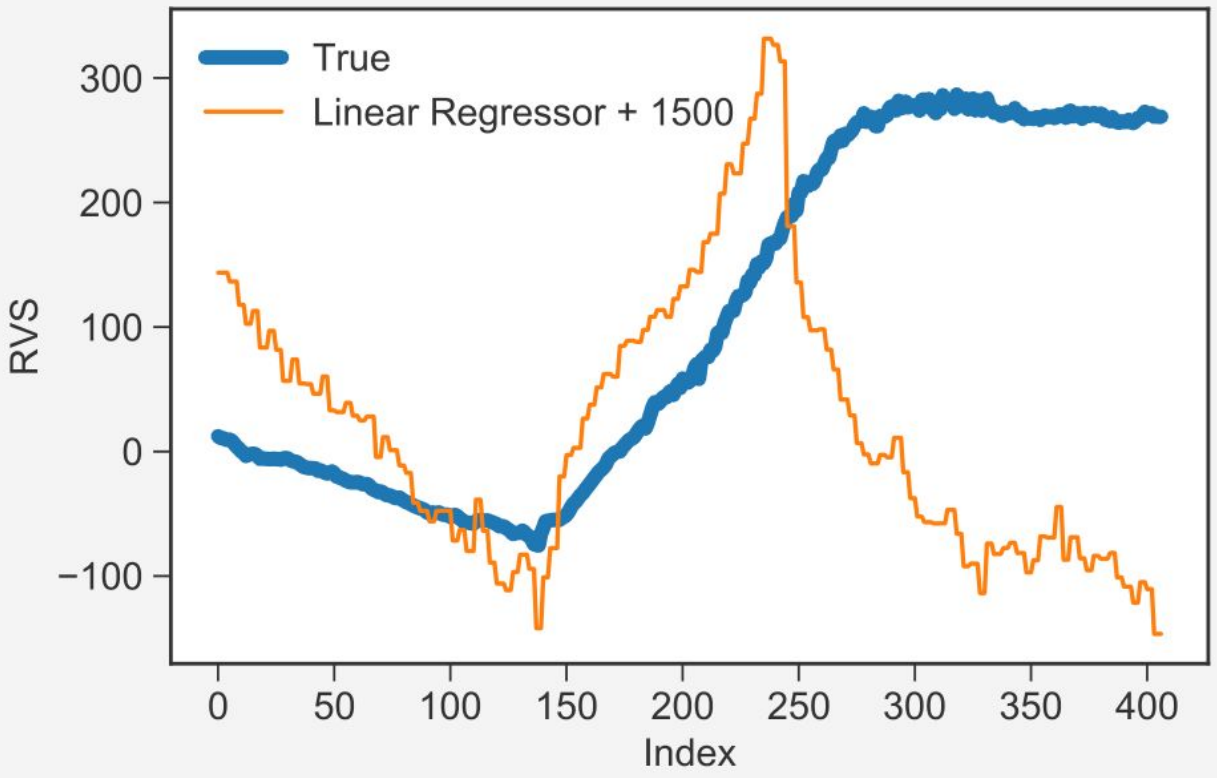
Predictions before scaling
LinReg on test 3/5



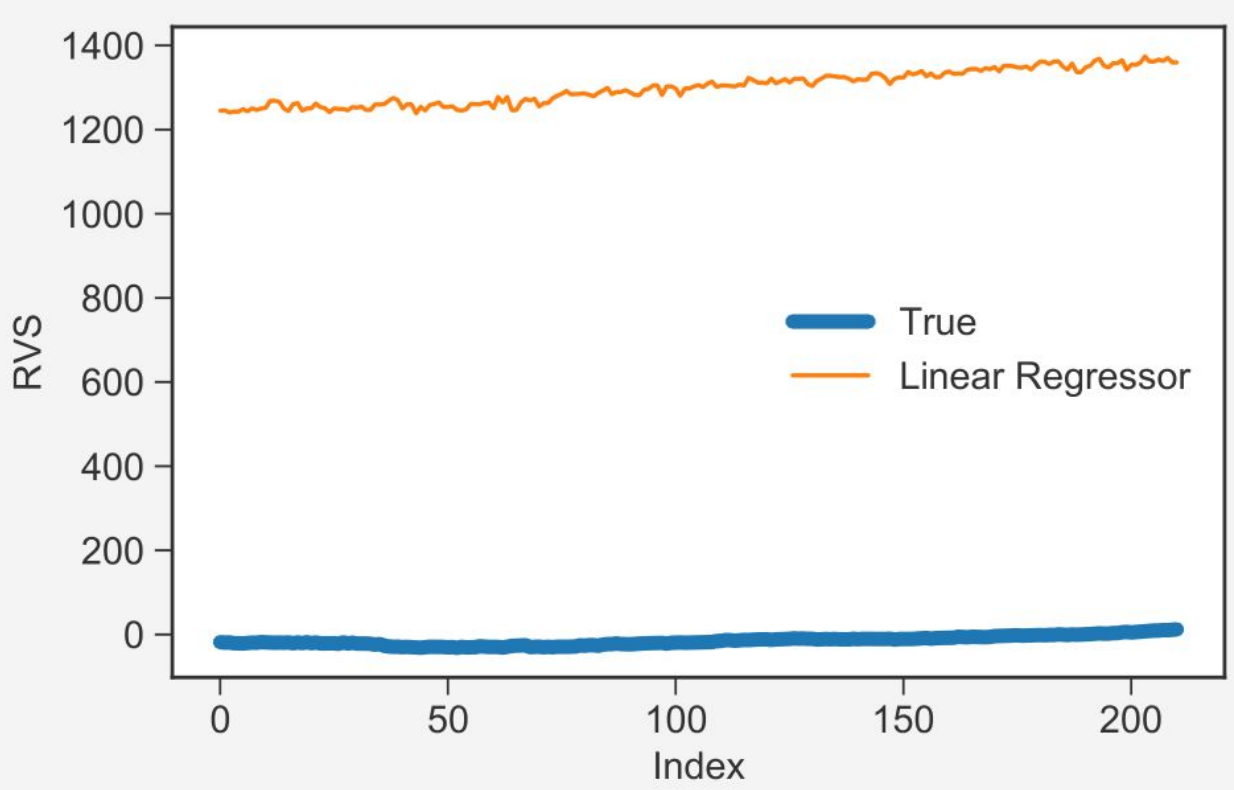


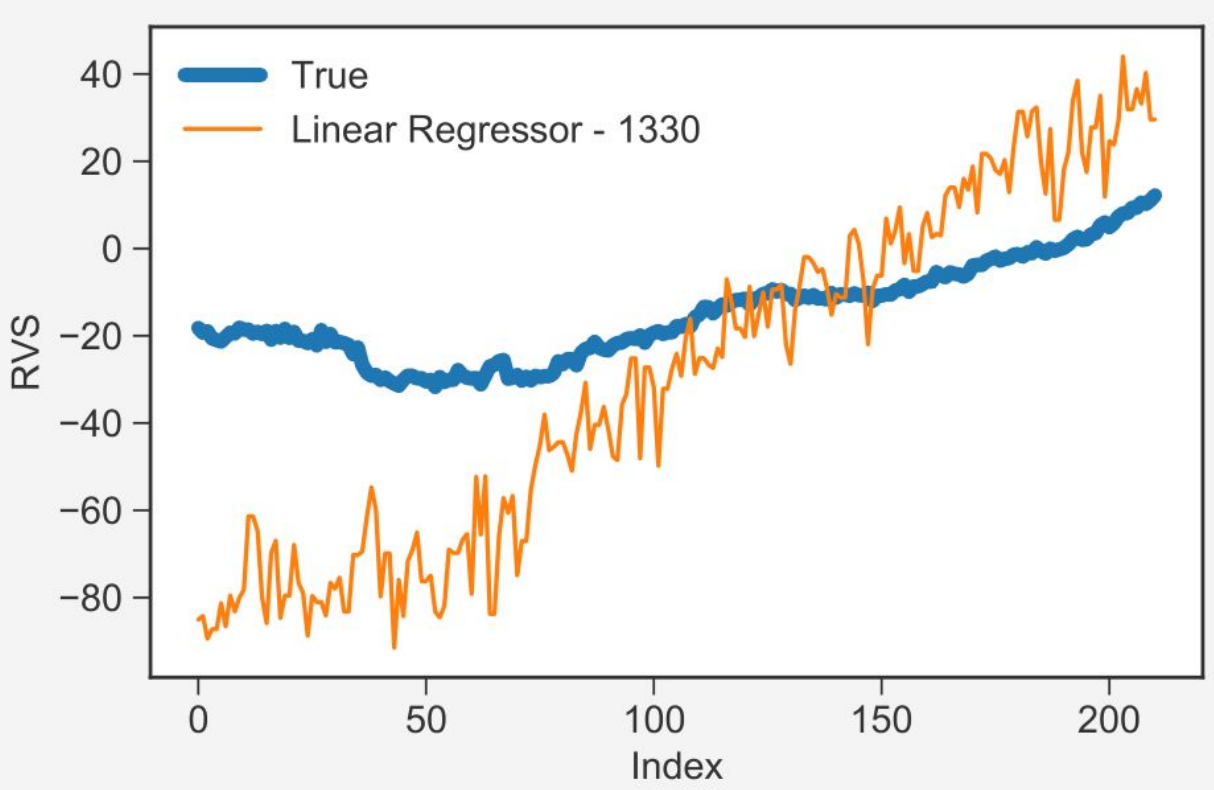
Predictions before scaling
LinReg on test 4/5





Predictions before scaling
LinReg on test 5/5





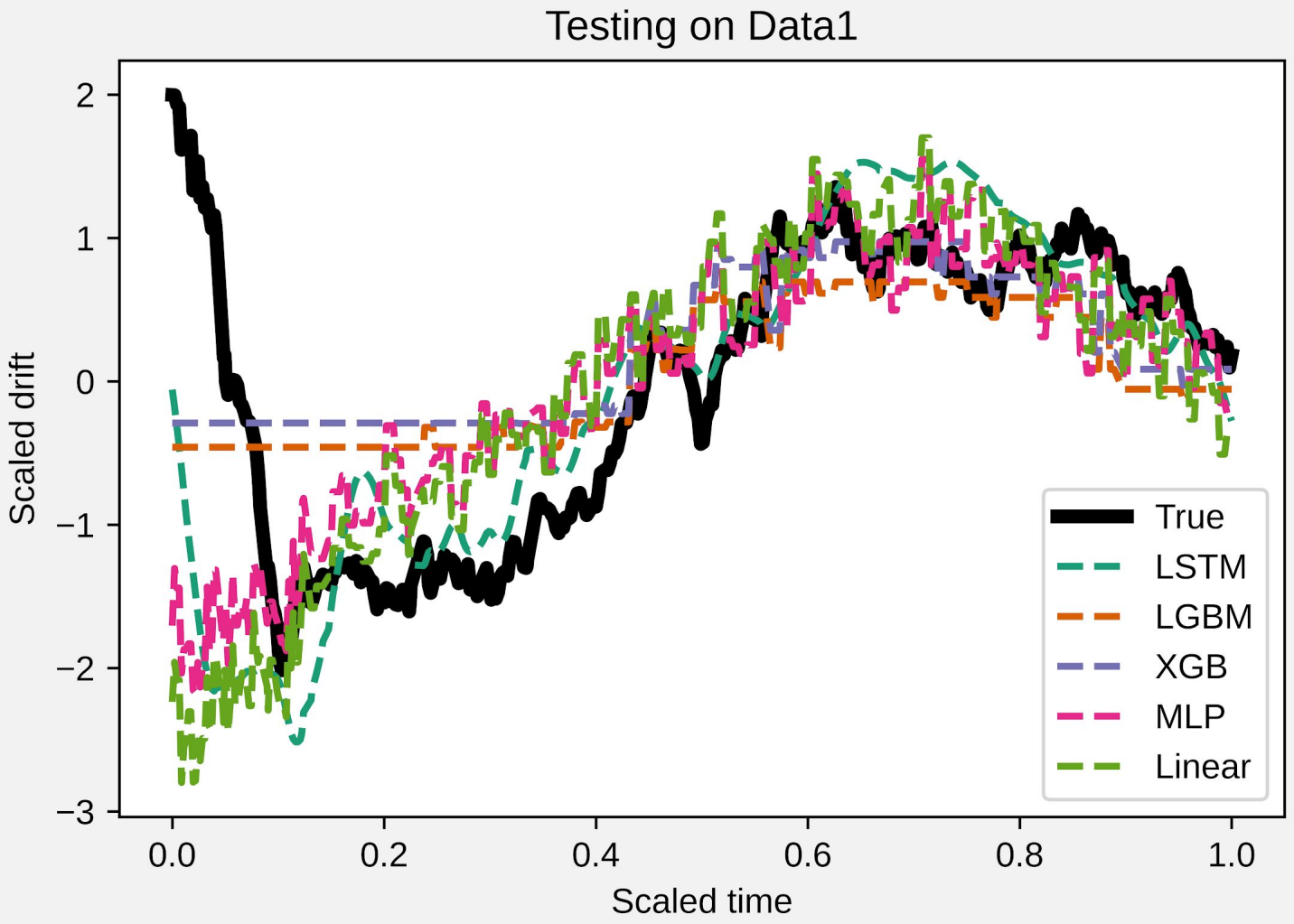
Optimization and results

The following slides present the results that were not presented during the presentation.

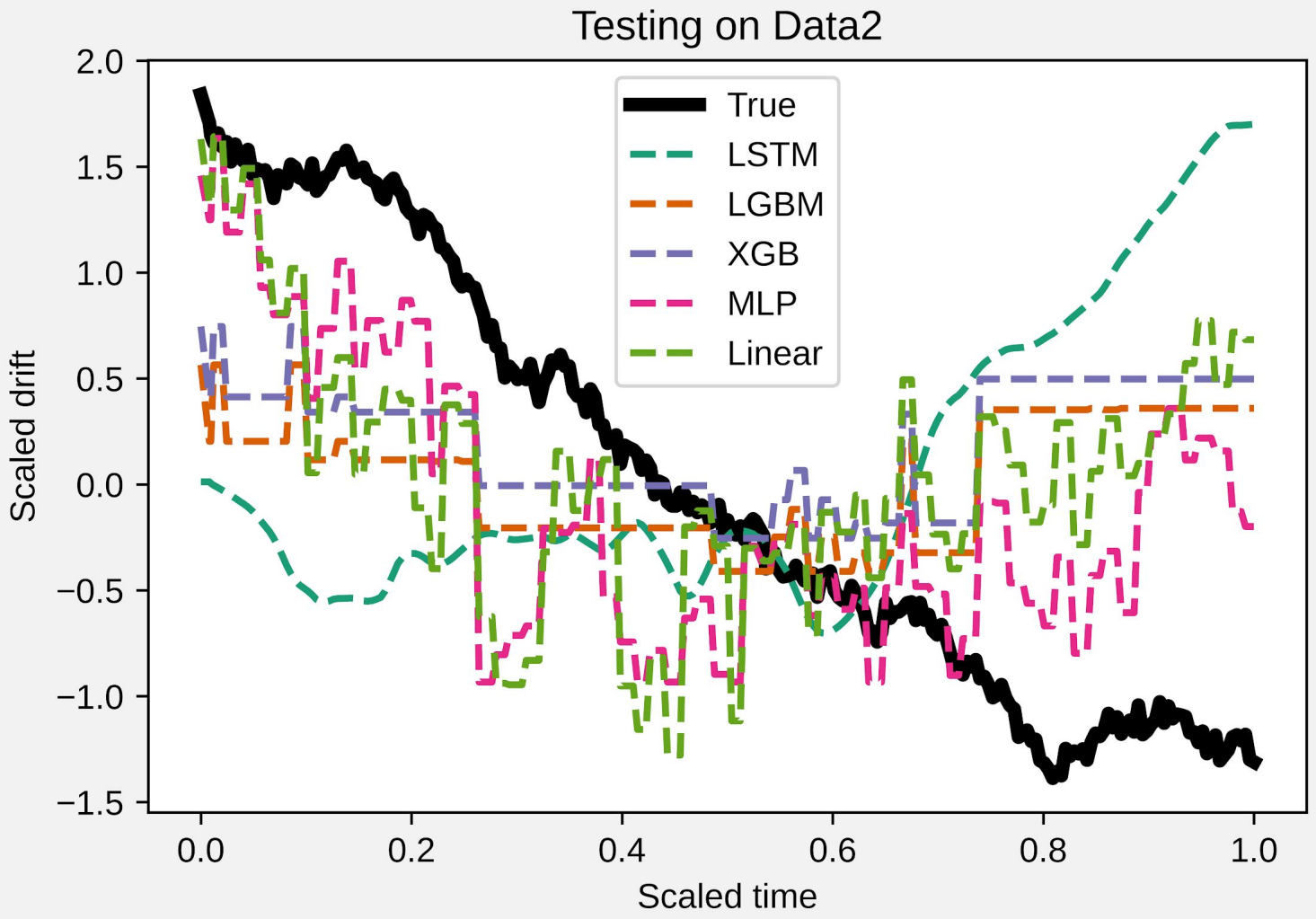
Each algorithm has been trained on all other dataset than the displayed one: Hence for the testing on dataset 1, we have been training on dataset 2, 3, 4, and 5.

To optimize the performance, we have exposed each model to hyperparameter optimization in the form of Bayesian Optimization with the cross validated estimate of the error as target.

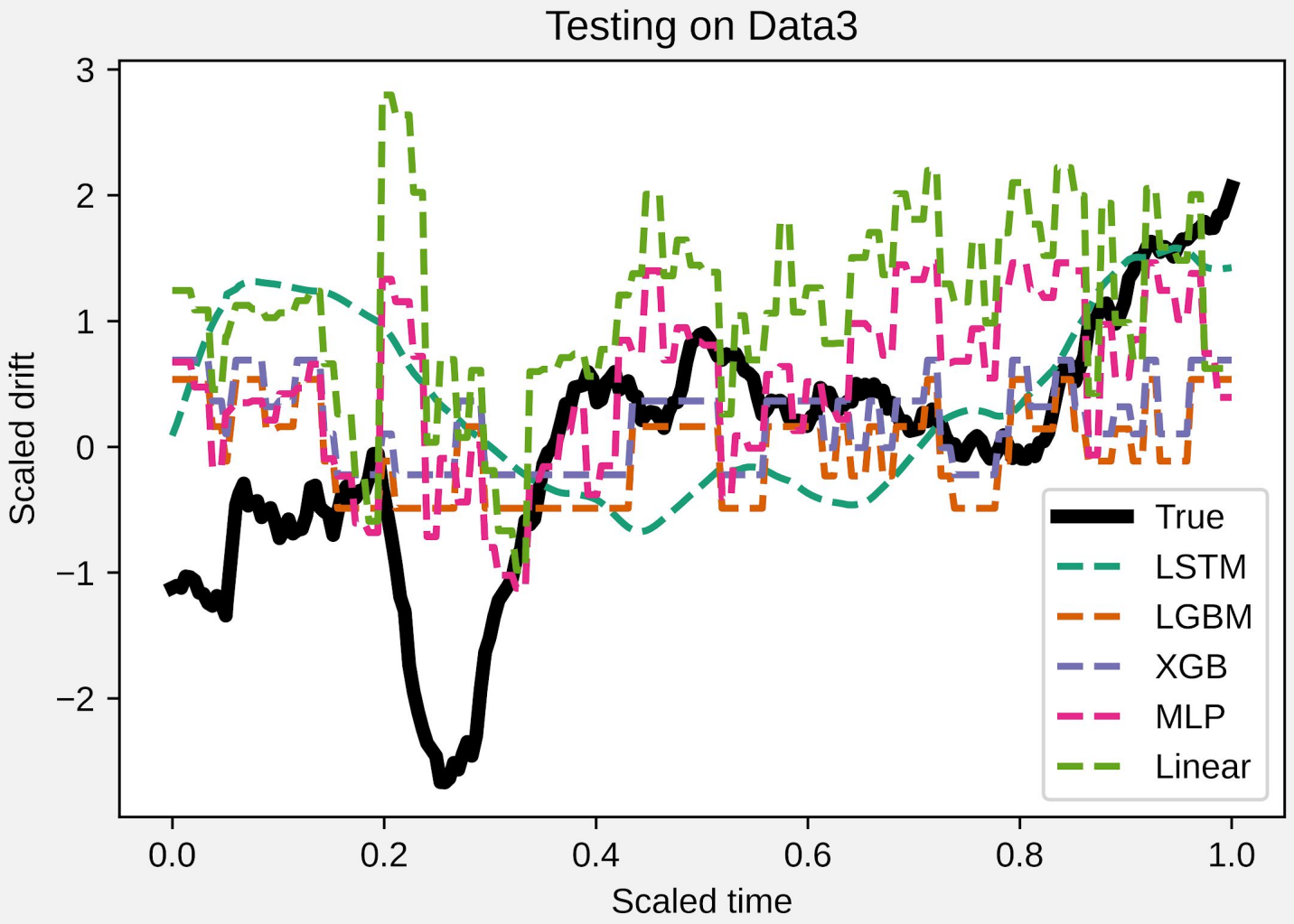
Results from testing on dataset 1



Results from testing on dataset 2



Results from testing on dataset 3



Results from testing on dataset 4