Prediction of the solar wind propagation delay for L1 to Earth using machine learning

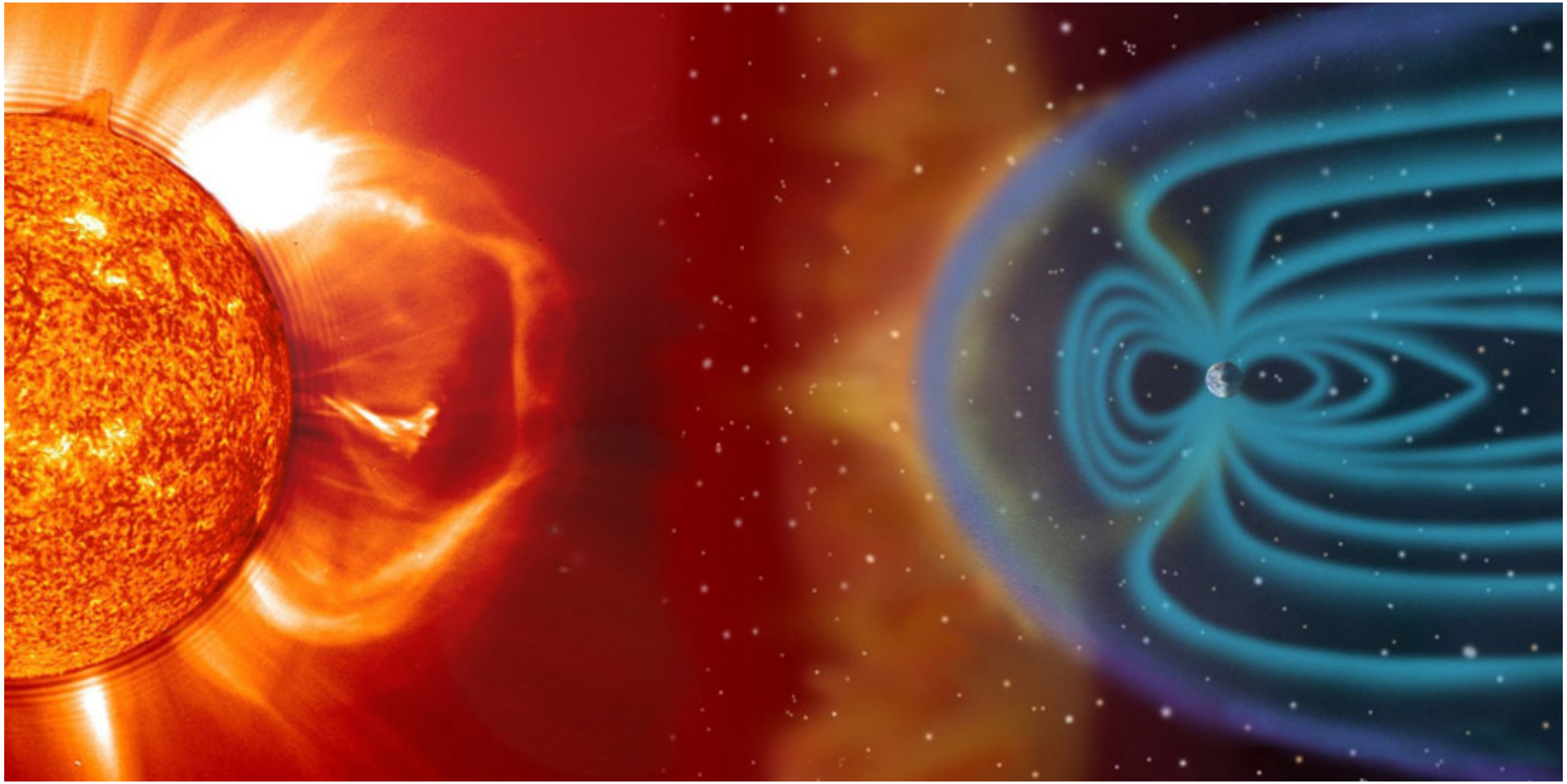
C. Baumann and A. E. McCloskey
Institute for Solar-Terrestrial Physics, Neustrelitz

DLR

Outline

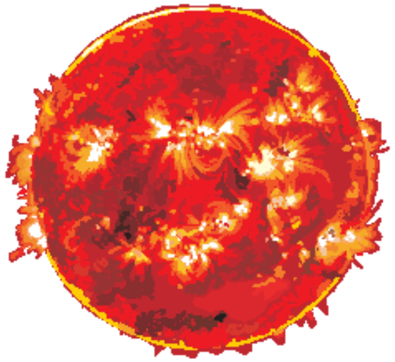
- Introduction
- Construction of database
- Comparison of ML with physical models of Solar wind Propagation delay
- Extract information from trained ML model
- Conclusion

Solar wind



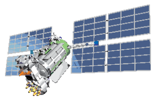
Introduction

Sun



0.99 AU

L1



0.01 AU
230 Re

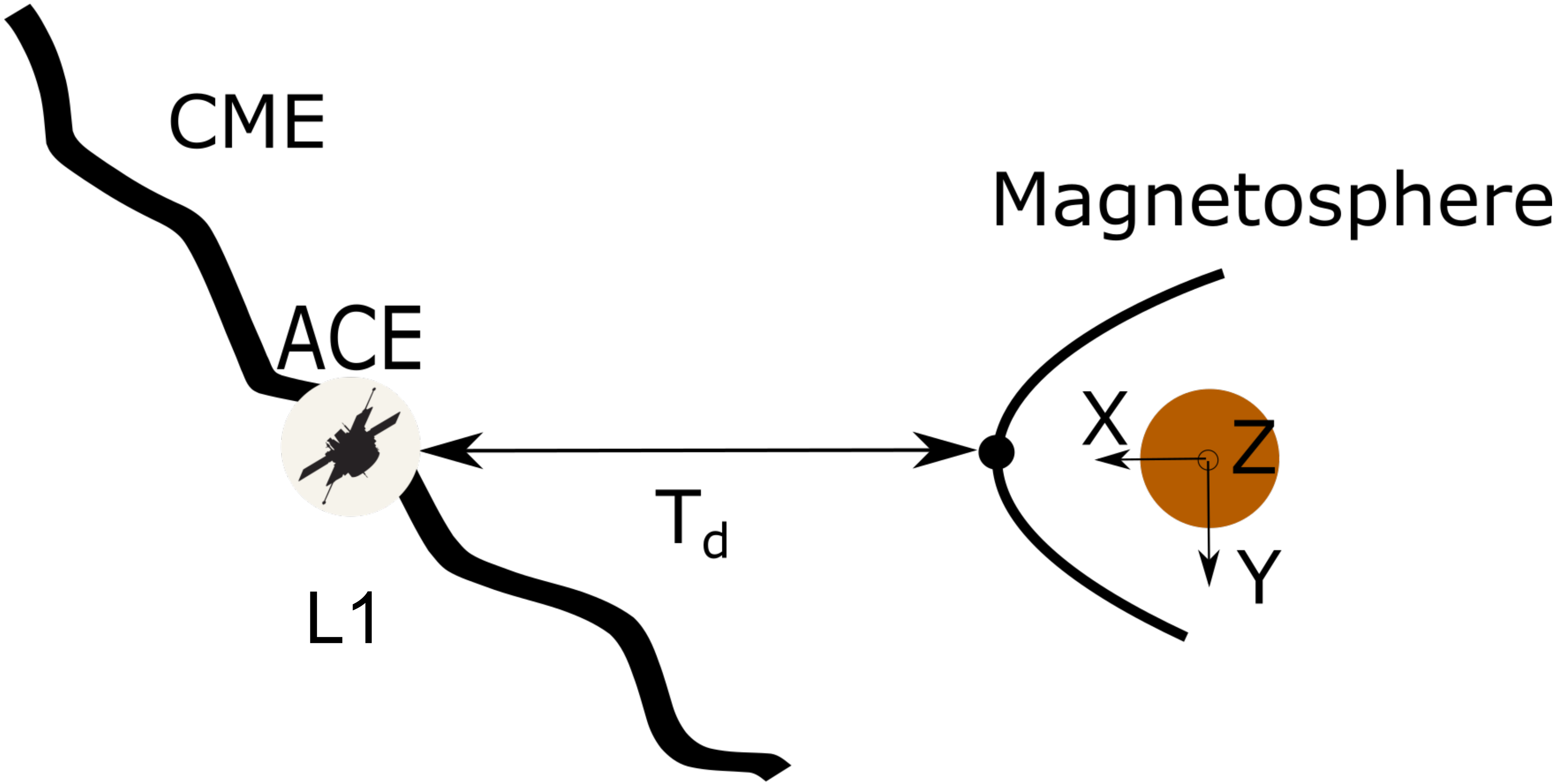
Earth

Orbit speed: 30km/s

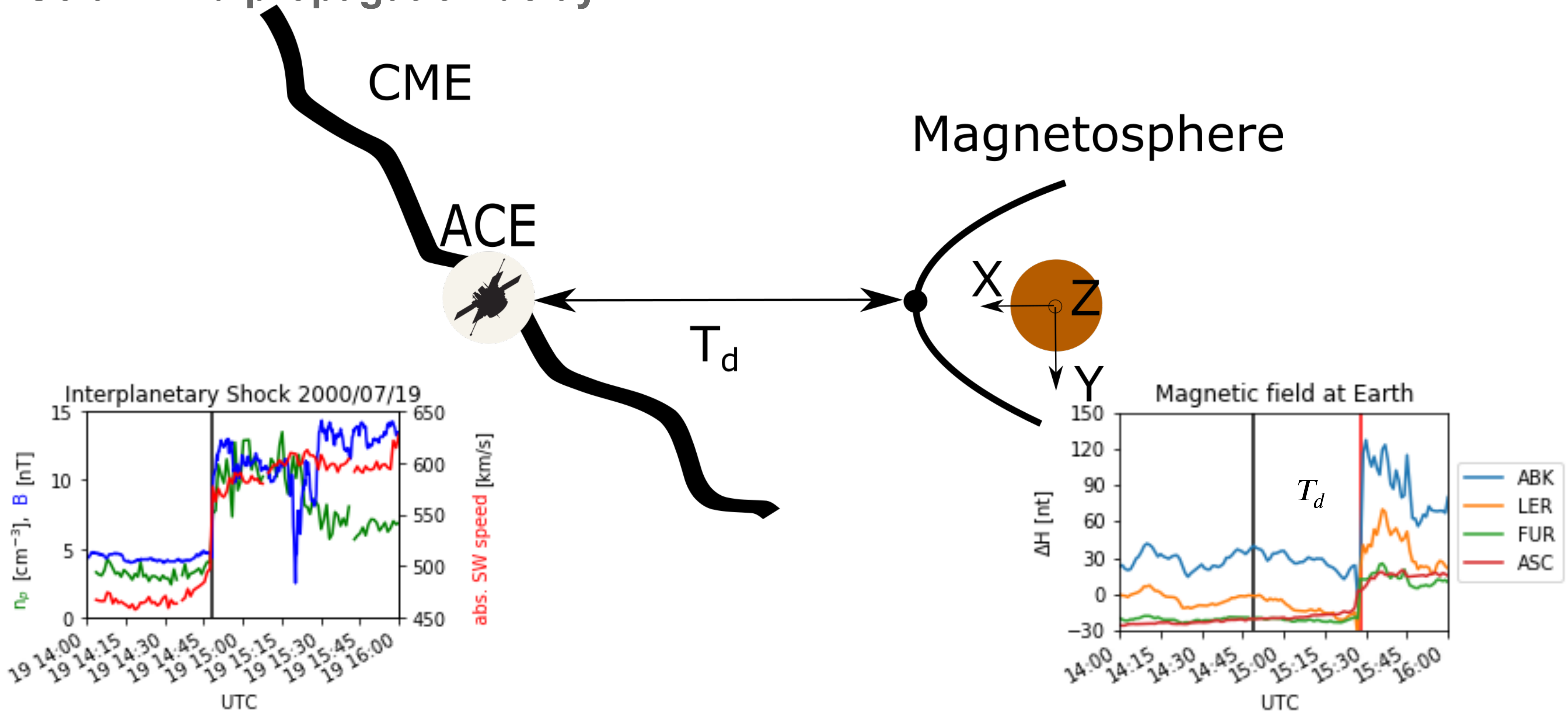
Solar wind T_d 1-3 d 15-80 min

SW Speed: 300-1000km/s

Solar wind propagation delay



Solar wind propagation delay



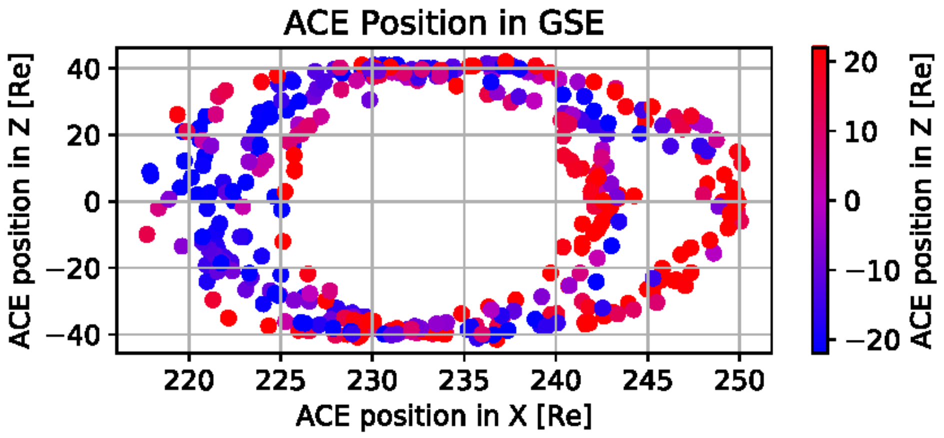
Database and ML approach

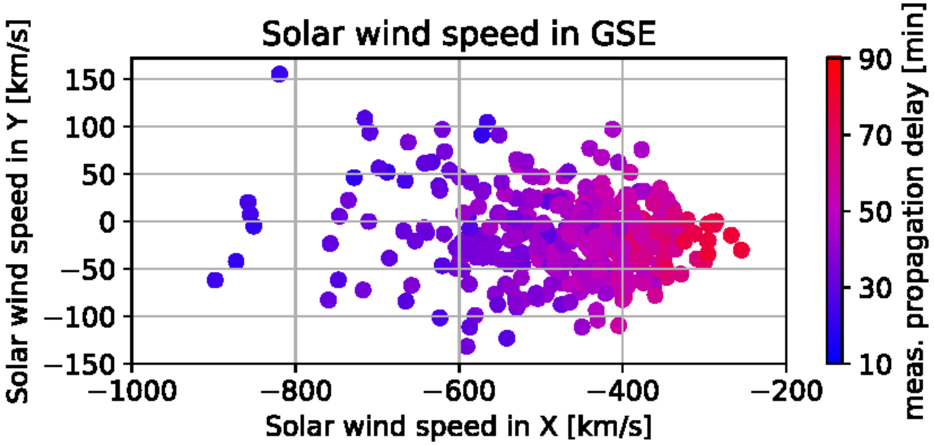
- Feature set:
- Solar Wind speed(V_x, V_y, V_z)
- Position of ACE (R_x, R_y, R_z)
- DST index, info on magnetospheric state

- Independent variable: T_d
- Solar wind Propagation delay

- Database contains 380 interplanetary shocks
- 380 individual measurements of T_d

- ML algorithms:
- Random Forest, Gradient boost, linear regression



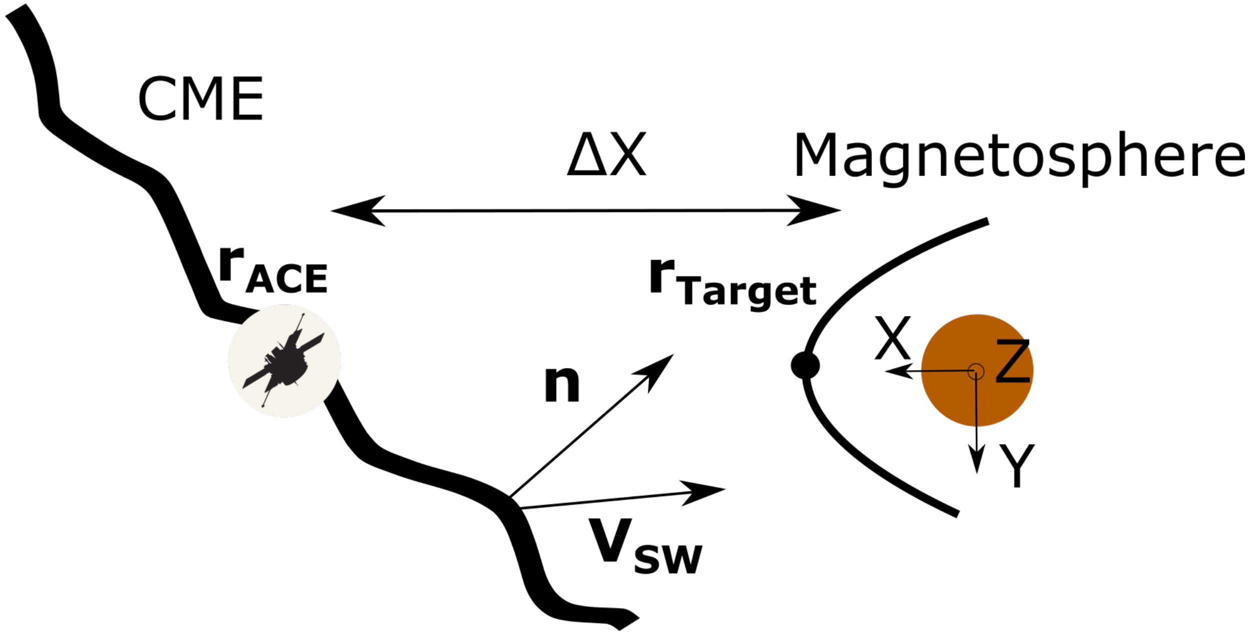


DLR

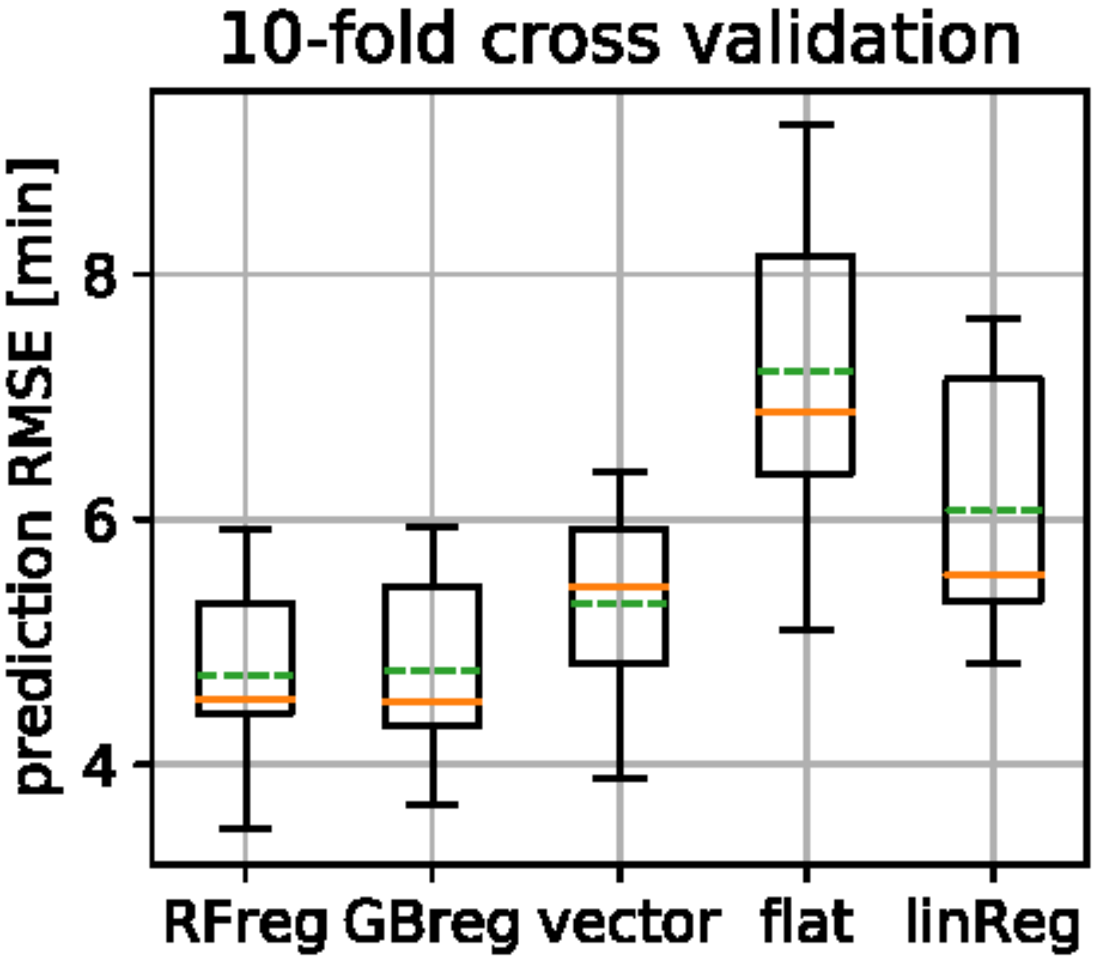
Physical models of the solar wind propagation delay

$$T_{flat} = \frac{\Delta X}{V_X}$$

$$T_{vector} = \frac{(\vec{r}_{target} - \vec{r}_{ACE}) \bullet \vec{n}}{\vec{V}_{SW} \bullet \vec{n}}$$

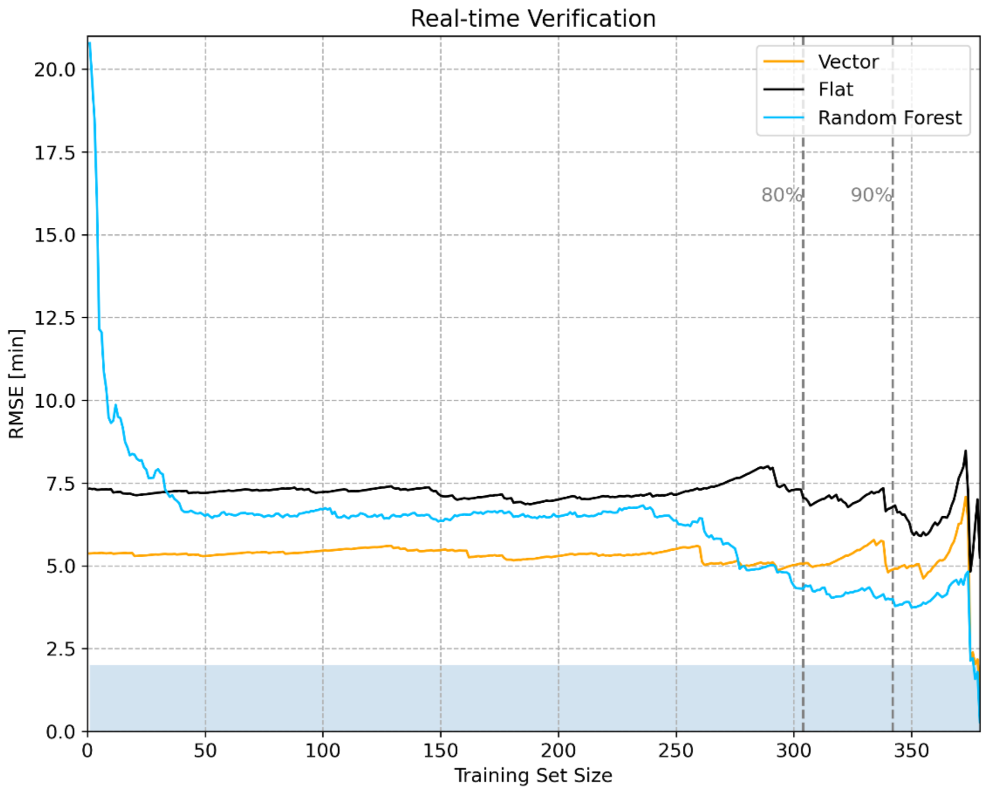


Comparison ML and physical model performance



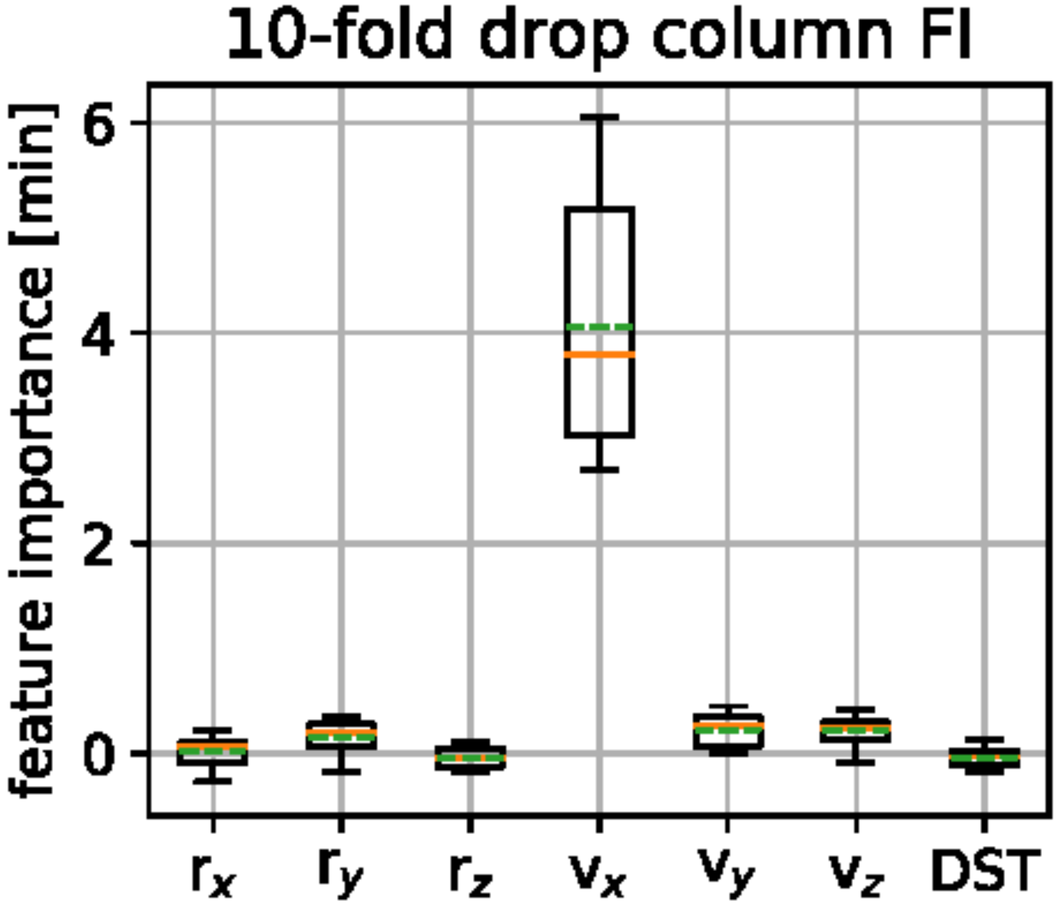
Real-time Validation

- Exploring model performance as a function of train/test split ratio on **unseen** data
- Vector method performs consistently well but RF **out-performs** when choosing **80/20** or **90/10** split
- All model performances **decrease & highly variable** as test size becomes **<10%** (not statistically reliable)
- RF model would benefit from more training instances



Feature importance

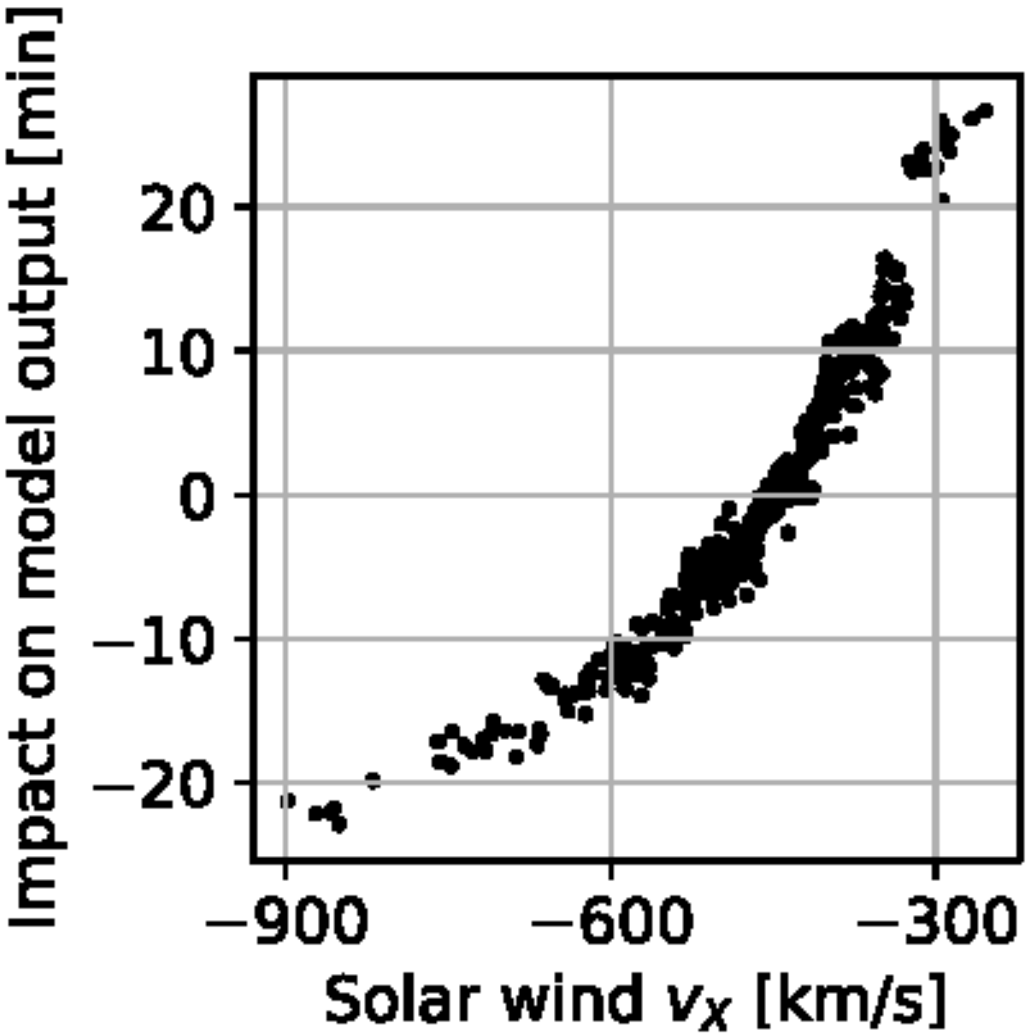
- Drop column feature importance compares a fully trained model with a model omitting a feature
- RMSE is used as metric
- Drop clumn FI: Change of RMSE when a feature is not used for training
- Positive values indicate worse performance, negative values indicate increase of performance
- Cross validation has been applied to investigate mean behavior



DLR

Shapley Value – Impact on model output

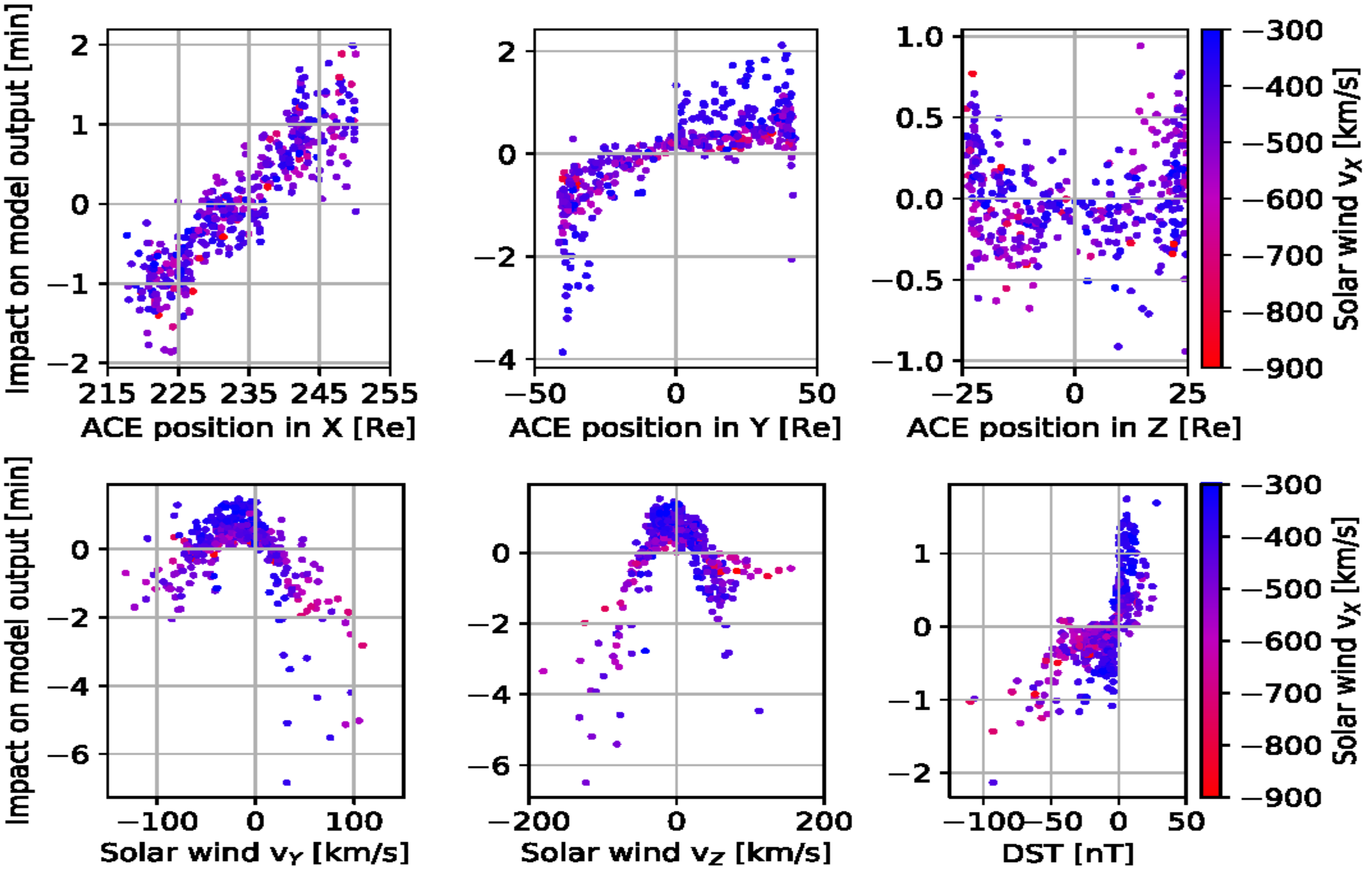
- Lloyd Shapley 1953 proposed a measure to identify the bonus due to cooperation within a cooperative game.
- The surplus that each player contribute to the outcome of the game is called Shapley value today.
- The principle can also be applied to the random forest regression of this study where its feature resemble Shapley's players.
- python package SHAP derives Shapley values



Mean random forest behavior

V_x[km/s]	V_y	V_z	R_x[RE]	R_y	R_z	DST	T_d[min]
-469	-14	0.6	233	0.89	0.29	-12	47

DLR

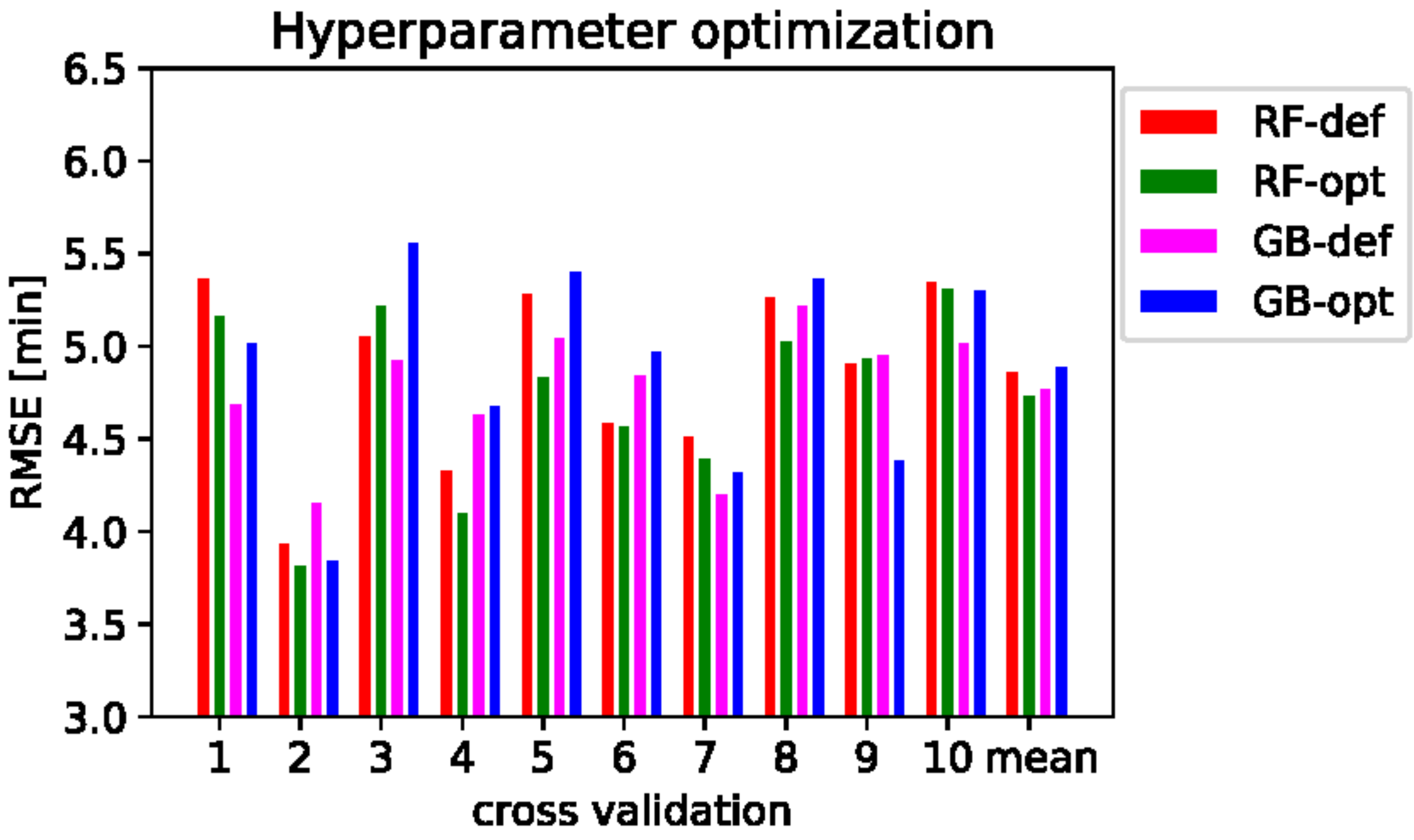

Impact on model output [min]
ACE position in X [Re]
ACE position in Y [Re]
ACE position in Z [Re]
Solar wind v_X [km/s]
Impact on model output [min]
Solar wind v_Y [km/s]
Solar wind v_Z [km/s]
DST [nT]
Solar wind v_X [km/s]

Conclusions

- Trained ML algorithm can predict the Solar wind propagation delay
- Results show better accuracy than flat and also vector method for SW delay prediction
- Shapley Value can be used to further analyse the analyze the
- The role of Earth‘s orbital speed within the SW delay problem has been discovered (maybe)
- Realtime application for L1 warning system

Thank you for your attention.

Extra


Hyperparameter optimization
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
RMSE [min]
1
2
3
4
5
6
7
8
9
10
mean
cross validation
RF-def
RF-opt
GB-def
GB-opt