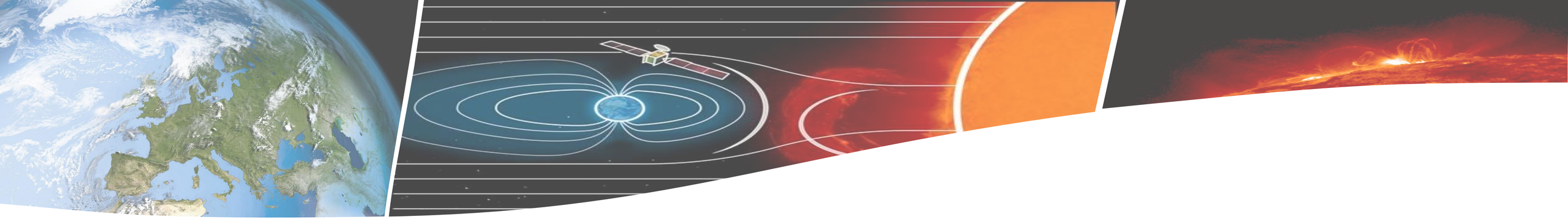


Application of Machine Learning Models in Predicting the Solar Wind Propagation from L1 monitors to the Earth's Bow Shock

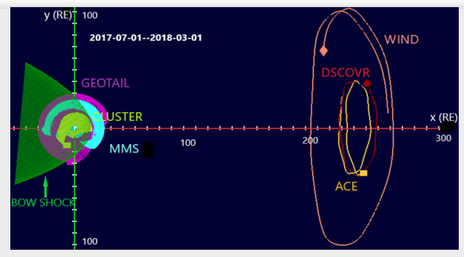
Samira Tasnim[1], Ying Zou[2], Claudia Borries[1], Carsten Baumann[3], Brian Walsh[4], Connor J. O'Brien[4], Sameer Gopali[5], and Huaming Zhang[5]

[1] Institute for Solar-Terrestrial Physics, DLR, Neustrelitz, [2]Johns Hopkins University Applied Physics Lab, Laurel, MD, USA, [3]WEMAG Netz GmbH, Schwerin, Germany,

[4] Department of Electrical and Computer Engineering, Boston University, [5]Computer Science Department, University of Alabama in Huntsville

Introduction

- The solar wind passing the Earth is an important driver of electrodynamic processes in the Earth's Magnetosphere-Ionosphere-Thermosphere (MIT) system
- Research and operational applications typically rely on measurements of solar wind monitors at the Lagrange point L1 as solar wind observation near Earth (at the bow shock) are very sparse
- The overarching goal of the project is to deliver machine learning models to specify and forecast near-Earth SW conditions based on spacecraft measurements around L1 by marrying the long history of multi-point SW measurements with the gradient boosting and random forest prediction models in the form of ensemble of decision trees
- We train the model to specify and/or predict the propagation time from L1 monitors to a given location upstream or at the bow shock.



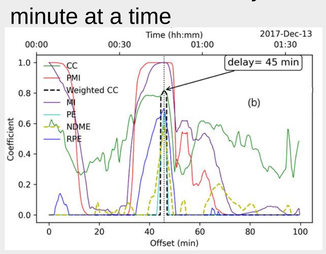
Orbits of ACE, WIND, DSCOVR, MMS, CLUSTER and GEOTAIL

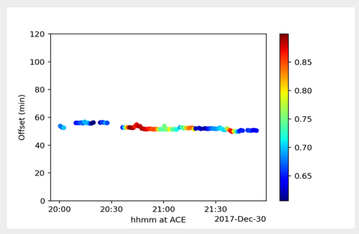
Methodology to Obtain SW Propagation Time

- To trace SW propagation, we perform the analysis on IMF clock angle:

$$\theta = tan^{-1}(B_y/B_z)$$

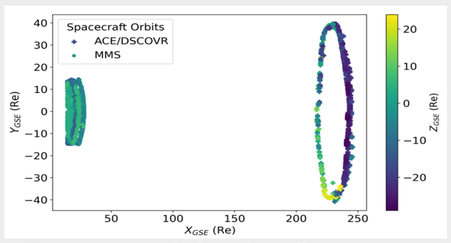
- We segment ACE in 20 minutes window and find the MMS data that best match the ACE data by sliding along 2 hours of data incrementing 1 minute at a time

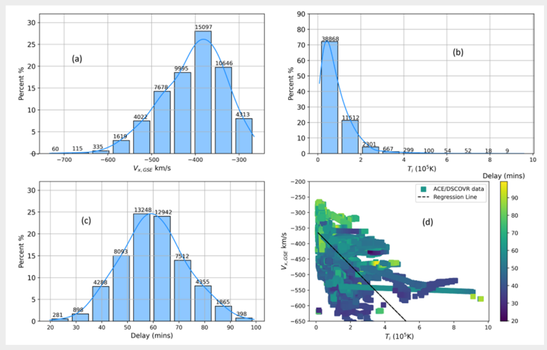


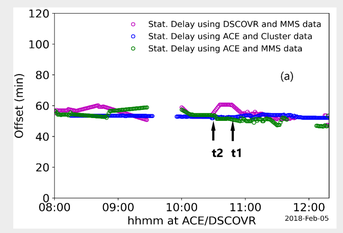


- To correlate IMF clock angle at L1 and near-Earth and obtaining propagation times, the algorithm computes
  1. Cross-correlation (CC) coefficient
  2. Plateau-shaped Magnitude Index (PMI)
  3. Dimensionless Measures of Average Error (NDME)
- Our analysis uses, Weighted CC = CC*PMI when max(CC) > 0.5 and NDME > 0.4

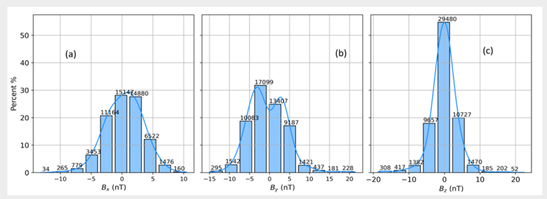
Data sets of input and target variables using multiple spacecraft pairs at L1 and near-Earth locations

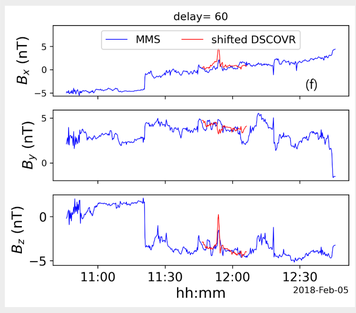






- The automated algorithm allows us to provide large sets of input and target variable using multiple spacecraft pairs at L1 and near-Earth location
- The algorithm facilitates easy access to data and the data sets can be used by anyone
- The developed algorithm generates a big dataset of 53880 events in the period from December 22, 2017, to April 30, 2024

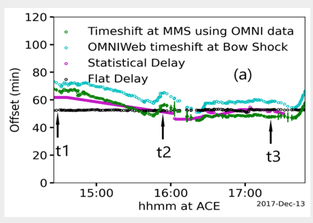


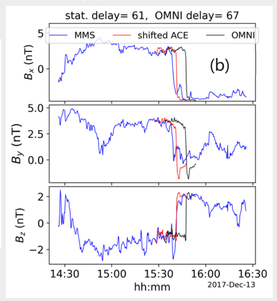


Statistical vs OMNIWeb Delay

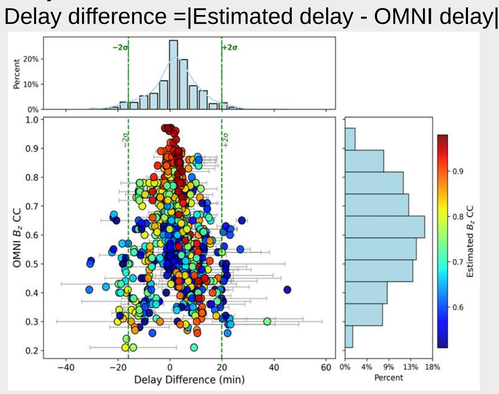
OMNIWeb Delay → Propagation delay calculated at a NE Monitor's location using OMNIWeb provided method and data

Stat Delay → Propagation delay estimated using our statistical approach/ correlation method

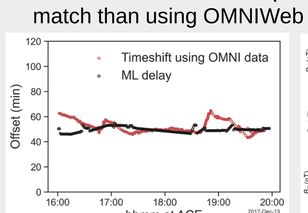


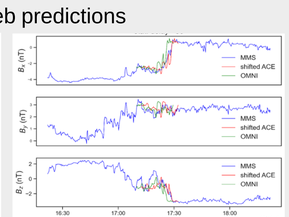


- Correlation coefficient values are calculated between B_z at L1 and shifted Bz using SW delays at the near-Earth Location
- B_z is shifted using OMNI delay and statistical delay
- Delay difference =|Estimated delay - OMNI delay|



- For selected cases, ML predictions shows better match than using OMNIWeb predictions





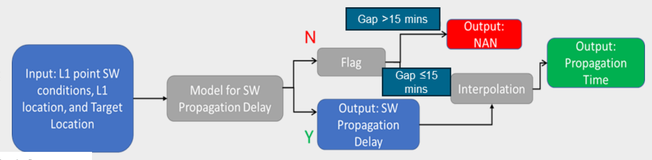
Application of Machine Learning Models in Predicting the Solar Wind Propagation from L1 to NE

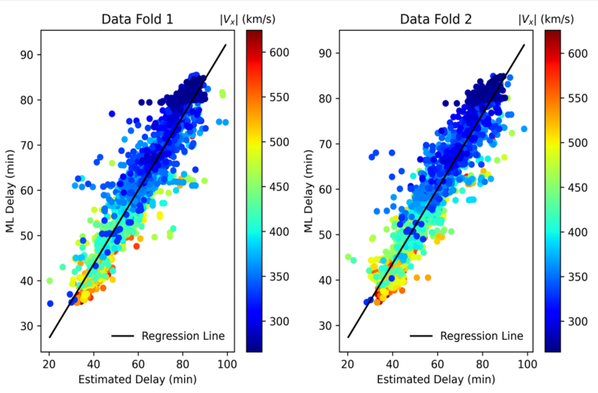
- We apply two ML models to predict SW propagation delay: i) Random Forest Regression (RF) and ii) Gradient Boosting (GB)
- GB and RF algorithms are applied together: a) To enable direct comparison between the RF and GB models and b) To quantify if the use of an ensemble-based ML model make a significant improvement to the overall performance
- The machine learning SW propagation delay can be described as

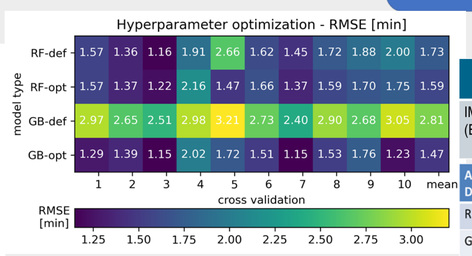
$$\Delta t_{ML} = f_{\mathcal{D}}(x)$$

Here $f_{\mathcal{D}}$ describes the ML algorithm trained on the data set $\mathcal{D}$ and x contains feature vectors

- We follow Baumann and McCloskey [2021]'s method, where we use Bayesian optimization based on the Gaussian process



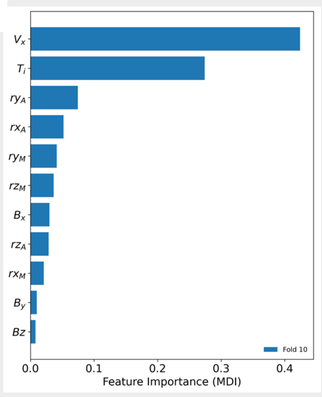


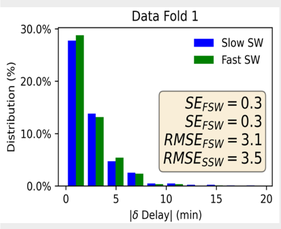


Input Feature Variables					Target Variable
IMF vector (B_x, B_y, B_z)	X component of SW velocity (V_x)	Ion Temperature (T_i)	Position of L1 monitor (x, y, z)	Position of NE monitor (x,y,z)	SW propagation delay

Algorithm Default/Optimized	Trees	Max features/Tree	Min samples split	Min sample leaf	Learning rate
Random Forest	150/800	9/3	2/20	1/20	NA
Gradient Boost	150/800	9/3	2/20	1/20	0.01

- The ML model predicted delay agrees well with the statistically estimated delay with an uncertainty of ± 5 minutes
- To optimize the hyperparameter and to assess the ML model performance, we employ a ten-fold cross validation approach





- Feature importance of the ML model and Correlations between feature vectors and the target SW delay are investigated. v_x has the highest feature ranking
- Validation: ML model predicted delays are compared with the results of physical models: 1) Flat Delay and 2) OMNI shifted delay using Phase Front Normal
- Evaluation of Deep Neural Network on test set results

Model	Training loss	Validation loss	RMSE	MAE	R^2	Adjusted R2
LSTM	0.9799	4.034	2.1978	1.2800	0.96831	0.96827
LSTM with dropout	1.5149	3.1631	1.926202	1.12128	0.97566	0.97562
MLP	4.8330	4.1961	2.24	1.094	09669	0.96688
MLP with dropout	23.1456	12.448	3.8065	2.5315	0.9049	0.9048

- Evaluation of Ensemble method on test set results

Model	RMSE	MAE	R2	Adjusted R2
Gradient Boosted Regression	1.7440	0.5058	0.9799	0.9798
Random Forest	1.3101	0.4981	0.9887	09886

- The ensemble methods outperformed the LSTM and MLP despite averaging out the time series information from the features
- The result can be explained by feature space being lower in dimension and training data being small as compared to other tasks such as language modeling where feature dimensions and amount of training data are very high and the LSTM model outperforms ensemble methods
- ML predictions provide similar results for he fast and slow solar wind

Summary and Conclusion

- The statistical approach conducts cross-correlation analysis to estimate SW propagation times and provides large sets of input and target variables
- We use multiple spacecraft pairs at L1 and near-Earth locations to train, validate, and test machine learning models
- The ML algorithm using these data sets helps to specify and predict (1) the propagation time from L1 monitors to a given location upstream or at the bow shock and (2) to forecast near-Earth SW conditions
- The obtained propagation times are then compared to OMNI. Factors that limit the OMNI accuracy are also examined
- The root mean squared error (RMSE) of RF is 1.3% and of GB is 1.7%, where the RMSE of MLP is 2.2%
- The ML model predicted delays are compared with the predictions of flat delays and OMNIWeb-provided delays. In the selected 100 cases, we found that about 10% of ML predictions result in a better match between the IMF features at the L1 point and those near-Earth, compared to the delay provided by OMNIWeb

References

i. Baumann, C., and McCloskey, A, Timing of the solar wind propagation delay between L1 and Earth based on machine learning, J. Space Weather Space Climate, 11, 2021

ii. Mailyan, B., Munteanu, C., and Haaland, S., What is the best method to calculate the solar wind propagation delay?, Annales Geophysicae, 26, 2008

iii. Case, N. A., and Wild, J. A., A statistical comparison of solar wind propagation delays derived from multispacecraft techniques, J. Geophys. Res., 117:A02101, 2012

iv. Tasnim, S., Zou, Y., Borries, C, Baumann, C, Walsh, B, O'Brien, C, Khanal, K., and Zhang, H., Estimation and Assessment of the Solar Wind Propagation Time from the Lagrange point L1 to the Earth's Bow Shock, Frontiers Astronomy and Space Science (Accepted)

Deutsches Zentrum für Luft- und Raumfahrt

Acknowledgments:
We acknowledge Deutsches Zentrum für Luft- und Raumfahrt (DLR), NASA grants 80NSSC20K1710 and 80NSSC21K0026 to support this research. We also acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb (or CDAWeb or ftp) service, and OMNI data.

Contact: Samira Tasnim
DLR, Institute for Solar-Terrestrial Physics
Kalkhorstweg 53, 17235 Neustrelitz
Tel.: +493981480316, Email: samira.tasnim@dlr.de