submitted to **Journal of Space Weather and Space Climate**

Timing of the solar wind propagation delay between L1 and Earth based on machine learning

C. Baumann[1] and A. E. McCloskey[1]

Deutsches Zentrum für Luft- und Raumfahrt, Institut für Solar-Terrestrische Physik, Kalkhorstweg 53, D-17235 Neustrelitz
e-mail: Carsten.Baumann@dlr.de

June 22, 2021

ABSTRACT

Erroneous GNSS positioning, failures in spacecraft operations and power outages due to geomagnetically induced currents are severe threats originating from space weather. Having knowledge of potential impacts on modern society in advance is key for many end-user applications. This covers not only the timing of severe geomagnetic storms but also predictions of substorm onsets at polar latitudes. In this study we aim at contributing to the timing problem of space weather impacts and propose a new method to predict the solar wind propagation delay between Lagrangian point L1 and the Earth based on machine learning, specifically decision tree models.

The propagation delay is measured from the identification of interplanetary discontinuities detected by the Advanced Composition Explorer (ACE) and their subsequent sudden commencements in the magnetosphere recorded by ground-based magnetometers. A database of the propagation delay has been constructed on this principle including 380 interplanetary shocks with data ranging from 1998 to 2018. The feature set of the machine learning approach consists of six features, namely the three components of each the solar wind speed and position of ACE around L1. The performance assessment of the machine learning model is examined on the basis of a 10-fold cross-validation.

The machine learning results are compared to physics-based models, i.e., the flat propagation delay and the more sophisticated method based on the normal vector of solar wind discontinuities (vector delay). After hyperparameter optimization, the trained gradient boosting (GB) model is the best machine learning model among the tested ones. The GB model achieves an RMSE of 4.5 min with respect to the measured solar wind propagation delay and also outperforms the physical flat and vector delay models by 50 % and 15 % respectively. To increase the confidence in the predictions of the trained GB model, we perform a operational validation, provide drop-column feature importance and analyse the feature impact on the model output with Shapley values.

The major advantage of the machine learning approach is its simplicity when it comes to its application. After training, values for the solar wind speed and spacecraft position from only one datapoint have to be fed into the algorithm for a good prediction.

Key words. solar wind propagation – ACE – machine learning

1. Introduction

Modern society is becoming increasingly vulnerable to space weather impacts. Orbiting satellites for communication and navigation, the once again emerging human space flight and power grids affected by induced currents require timely information on imminent severe space weather events. One of the main drivers of space weather at Earth is the continuous flow of solar wind. However, the nature of the solar wind is variable and ranges from a slight breeze of electrons and protons, to fast storms of energetic particles containing ions as heavy as iron. For the surveillance of the solar wind, several spacecraft have been installed at the Lagrangian point L1. The ACE spacecraft has been, and still is, a backbone for early warnings of severe solar wind conditions (Stone et al., 1998).

For precise forecasts of the ionospheric and thermospheric state, the expected arrival time of these severe solar wind conditions at Earth's magnetosphere is needed. For that purpose, modelling the propagation delay of the solar wind from spacecraft at L1 to Earth has been a long-standing field of research (e.g. Ridley, 2000; Wu et al., 2005; Mailyan et al., 2008; Pulkkinen and Rastätter, 2009; Haaland et al., 2010; Cash et al., 2016; Cameron and Jackel, 2016). In particular, communication and navigation service users are interested in timely and reliable information about whether to expect a service malfunction or outage at a specific upcoming moment. On the other hand, research topics such as the timing of the onset of polar substorms (e.g. Baker et al., 2002) also benefit from precise information when a potential triggering solar wind feature reaches the magnetosphere.

The above mentioned techniques for the prediction of the solar wind propagation delay depend on the presence of a shock in the interplanetary medium and provide a velocity-based time delay. Additional approaches to propagating the solar wind include hydrodynamic modeling (e.g. Kömle et al., 1986; Haiducek et al., 2017; Cameron and Jackel, 2019), which model the physical evolution of the solar wind plasma as it travels to Earth.

The measurement of the solar wind propagation delay is usually done by identifying its distinct features at spacecraft around L1 and Earth orbiting satellites which temporarily probe the solar wind directly (CLUSTER, MMS, Van Allen Probes). These features can be turnings of the interplanetary magnetic field (IMF) (Ridley, 2000) and even discontinuities in the solar wind caused by Coronal Mass Ejections (CMEs) or Corotating Interaction Regions (CIR) (e.g Mailyan et al., 2008) which are used in this study as well. The magnetosphere on the other hand is also suited to serve as detector of solar wind features. Magnetometer stations observe the state of Earth's magnetic field on a continuous basis. As CMEs and CIRs pass the Earth, they can cause significant disturbance to the magnetosphere, leading to so-called sudden commencements in the magnetic field (Gosling et al., 1967; Curto et al., 2007). These sudden commencements are detected by ground-based magnetometers across the globe (Araki, 1977), allowing for the timing of the solar wind propagation delay just as well as space-based magnetometers.

This study compiles a database of the solar wind propagation delay based on interplanetary shocks detected at ACE and their sudden commencements (SC) at Earth. The database consists of timestamps of the shock detections at ACE and the following SC detections by groundbased magnetometers. The study by Cash et al. (2016) used the same principle to measure the propagation delay. The database serves not only as a basis to assess the performance of the physical models of the SW propagation between L1 and Earth, but also as a training set for a novel approach based on machine learning (ML).

NASA's well known OMNI database (https://omniweb.gsfc.nasa.gov/) applies the method of Weimer and King (2008) to provide solar wind propagation delays (i.e. the so called timeshift) for 30 R_E ahead of Earth continuously. Cash et al. (2016) tested the ability of Weimer's method in a realtime application and found that it suffers from caveats introduced by additional assumptions applied to the initially shock based method in order to work with continuous data as well.

This study introduces a machine learning method to predict the solar wind propagation delay. The training dataset is defined in a way that only one datapoint of L1 spacecraft data is needed for input, enhancing its flexibility for the use of continuous data as well and may also enable a potential realtime application in the future. However, as the database for training is comprised of CME and CIR cases only, the valid generalization to a continuous application of the ML approach remains unresolved. The present work can be seen as a first proof of concept that machine learning is indeed able to predict the solar wind propagation delay.

In recent years there has been an ever-increasing number of studies in the field of space weather that have made use of ML algorithms. More specifically, these ML algorithms have been particularly successful for the purpose of prediction, including the prediction of CME arrival times based on images of the Sun (Liu et al., 2018), solar wind properties (Yang et al., 2018), geomagnetic indices (Zhelavskaya et al., 2019) and even predictive classification of (storm) sudden commencements from solar wind data (Smith et al., 2020). For an overview on ML applications for space weather purposes we recommend the review by Camporeale (2019). The advantage of using an ML-based approach, instead of a solely empirical or physics-based model, is that ML models don't require as many a priori assumptions and are generally less computationally intensive.

The present study investigates the possibility to use ML to predict the solar wind propagation delay and is structured as follows. Section 2 describes the measurement technique and database of the solar wind propagation delay used in this study. Section 3 introduces the physical models of SW propagation delay and also the new machine learning approach. Section 4 shows the results of the model comparison and an analysis of the trained ML algorithm. The discussion of the results is carried out in Sec. 5. In Sec. 6 we draw the conclusions from our findings.

2. Delay measurement and database

The following section presents the methods of how the solar wind propagation delay has been measured and gives an overview of the contents of the database. The database of this study is solely comprised of ACE observations of interplanetary shocks and their subsequent sudden commencements at Earth for the years 1998 until 2018.

The database consists of 380 cases that have been identified by ACE and magnetometers on the Earth's surface. The used ACE level 2 data has been provided by the ACE Science Center at Caltech and consists of a timeseries with 64 s time resolution. The times of interplanetary shocks have been identified from shock lists (Jian et al., 2006; Oliveira and Raeder, 2015) and the website ipshocks.fi maintained by University of Helsinki. These lists combine ACE, Wind and DSCOVR detections of interplanetary shocks and do not always list data for all three spacecraft. In order to increase the number of shock detections for ACE, we have searched ACE data around times that listed shocks for Wind or DSCOVR but not for ACE. In case ACE did detect the shocks as well, these ACE detections are then added to the database used in this study. The most recent

interplanetary shocks (post 2016) have been identified by visual inspection of ACE data during high geomagnetic activity. The authors cannot guarantee the capture of all interplanetary shocks from ACE data within the database presented here.

Figure 1 (top) shows a typical interplanetary shock as it is measured by the ACE SWEPAM (McComas et al., 1998) and MAG (Smith et al., 1998) instruments on the 19th July 2000 at 14:48 UTC. The solar wind propagation delay for this case is identified from the sudden commencement that this interplanetary shock causes in Earth's magnetosphere (Fig. 1 bottom panel). The term sudden commencement describes a magnetospheric phenomena which ground based magnetometers can observe when the magnetosphere is compressed by the impact of an interplanetary shock, i.e. due to the sudden change of the solar wind dynamic pressure (e.g. Curto et al., 2007). The signature of a sudden commencement is defined as a sudden change of the horizontal component of the magnetic field (ΔH). ΔH is the difference between the actual horizontal component H and a baseline (ΔH=H-H_0).

The sudden commencement is identified from magnetometer data at different locations from high latitudes down to equatorial regions. The search for the sudden commencements were restricted to times 0.25 to 1.5 h after the detection of an interplanetary shock at ACE. An additional constraint of the identification of the sudden commencement is that it happens quasi-simultaneously (<1 min) at all latitudes (e.g. Engebretson et al., 1999; Segarra et al., 2015). In our study we have used magnetometer stations at Abisko (ABK, 68°N), Lerwick (LER, 60°N), Fürstenfeldbruck (FUR, 48°N), Bangui (BNG, 4°N), Ascension Island (ASC, 8°S), and Mawson (MAW, 67°S) which are part of the INTERMAGNET consortium (Love and Chulliat, 2013). The identification of sudden commencements is based on 1 min magnetometer data and has been done using the SuperMAG service at JHAPL (Gjerloev, 2012). For this analysis, we use the northward component of the magnetic field given by SuperMag to identify the times of sudden commencements.

In 95 % of the cases the identification was unambiguous with the above stations. However, the identification of sudden commencements during active geomagnetic conditions is not possible at high latitudes. That is because the already disturbed magnetic field prevents a clear identification of a sudden commencement. During ≈5 % of the cases, additional magnetometer stations Tamanrasset (TAM, 22°N) and Toledo (TOL, 40°N) have been used to detect the sudden commencement without ambiguities. An ACE detection/sudden commencement pair was added to the database only in the case of its simultaneous identification at 5 different stations. Weak interplanetary shocks detected at ACE, that did not cause a geomagnetic sudden commencement, do not allow a measurement of the solar wind propagation delay and are discarded from the database.

The moment of the interplanetary shock at ACE (T_{ACE}) has been set to the datapoint when the solar wind speed reaches its downstream (high) value (black dashed vertical line in Fig. 1). So the database consists of just 380 datapoints at the individual time T_{ACE}, no timeseries data before the shock is used for the ML propagation delay approach. The moment of the sudden commencement (T_{SC}) at Earth's magnetosphere is set to the sudden increase of δH. The propagation delay (T_D) is defined as the difference between both times.

$$T_D = T_{SC} - T_{ACE} \tag{1}$$

The systematic error of the measured solar wind propagation delay is at least 2 min, because the time resolution of ACE SWEPAM data and of the magnetometer data is 1 minute each.

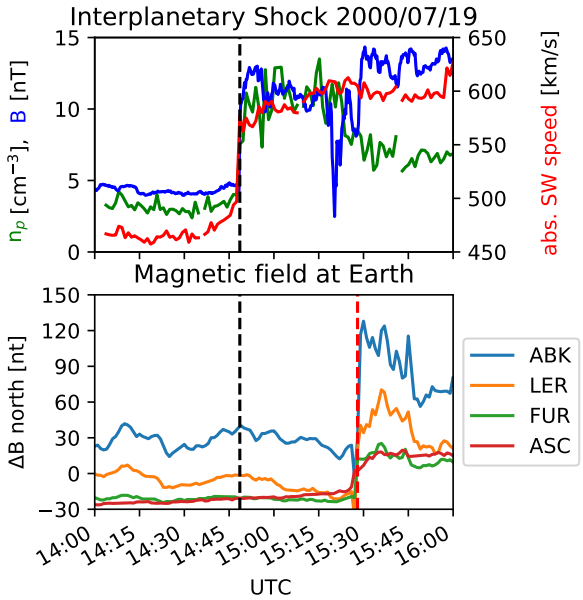

Fig. 1. Top panel shows timeseries of proton density n_p, magnetic field B and absolute solar wind speed measured by ACE, bottom panel shows magnetic northward component residuals for magnetometer stations Abisko, Lerwick, Furstenfeldbruck, and Ascension Island, black and red vertical dashed line indicate the interplanetary shock at ACE time 2000/07/19 14:48:35 and a sudden commencement in Earth's magnetosphere at 15:27:00 resulting in a delay of 38 min.

Figure 2 shows the database for the measured propagation delays and additional parameters that have been extracted from the ACE data at the 380 individual times, T_{ACE}. The top panel shows the position of the ACE satellite at the time of the IP shock detection. It is evident, that the shown positions of ACE represents its Lissajous orbit around Lagrange point L1. The position of ACE is important for the calculation of the propagation delay and is used for the physical models and the statistical ML model as well. The bottom panel shows solar wind speed in X and Y-direction measured at T_{ACE} color coded with the measured propagation delay T_D. The propagation delay varies between 20 min for extremely fast ICMEs around 1000 km/s and nearly 90 min for slow shocks around 300 km/s. The solar wind speeds and the position of ACE are given in GSE coordinates, therefore higher solar wind speeds show larger negative values in X-direction. Not shown in Fig. 2 is the solar wind speed in Z-direction, which ranges between -150 and 150 km/s. So the database consists of just 380 datapoints at the individual time T_{ACE}, no timeseries data before the shock is used for the ML propagation delay approach.

The database described here has been made available on Zenodo (Baumann and McCloskey, 2020). It contains the times T_{ACE} and T_{SI} for all 380 interplanetary shocks and sudden commencements respectively. It also contains the ACE measurement of all three components of the solar wind speed (v_x, v_y, v_z) and the position of ACE (r_x, r_y, r_z) at T_{ACE}.

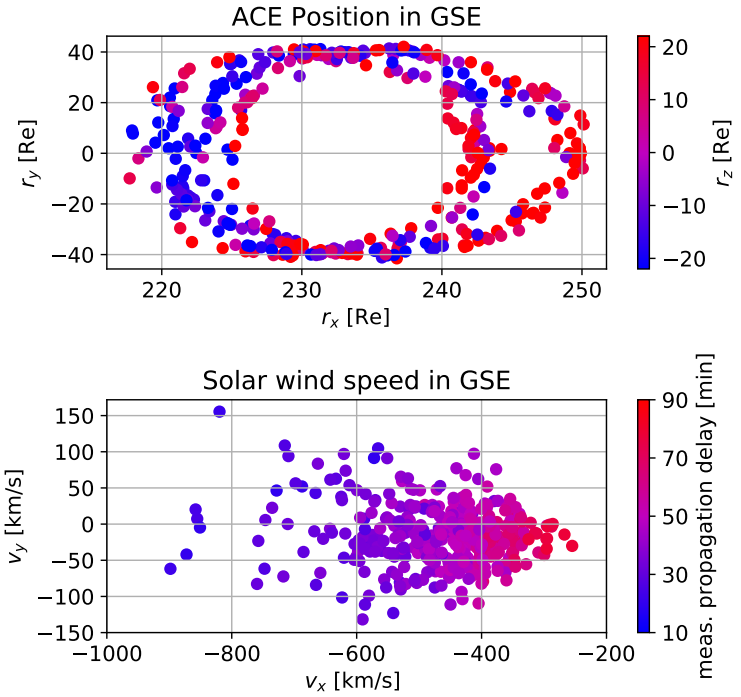

Fig. 2. Overview of the database for the calculation of the solar wind propagation delay, top panel shows the ACE position in X, Y, Z, bottom panel shows the measured solar wind propagation delay and the solar wind speed in X, Y, all units are given in GSE coordinates, except the propagation delay [min].

3. Propagation delay models

The solar wind propagation delay between L1/ACE and Earth can be divided into three parts as indicated in Fig. 3. Firstly, the delay between ACE and Earth's bow shock, i.e. where the solar wind speed drops significantly. Secondly, the time between the impact at the bow shock and magnetopause. Thirdly, the delay between the impact at the magnetopause and the start of space weather effects on the ground, e.g. geomagnetically induced currents. While the first part of the propagation delay can be seen as a pure convection of the solar wind (e.g. Mailyan et al., 2008), the other two parts of the delay depend on the geomagnetic conditions and the type of incoming solar wind feature as well as its characteristics. This study focuses on the delay from the Lagrangian point L1 to the magnetopause (see Sec. 2). Timing biases contained within the derived propagation models could account for the differences in timing delay output for the three types of models that will be introduced later in this section.

Otherwise, the solar wind propagation delay is usually addressed by measuring the delay between spacecraft, i.e. a monitoring spacecraft at L1 and an Earth satellite just outside of the terrestrial magnetosphere. The launch of Wind and ACE for the purpose of solar wind monitoring at the L1 point initiated multiple studies on the physical modelling of the solar wind propagation delay (Ridley, 2000; Horbury et al., 2001; Weimer et al., 2003; Mailyan et al., 2008). It has to be noted that the solar wind monitors at L1 are not perfect measures of the solar wind that will eventually interact with Earth's magnetosphere at a later time (e.g. Borovsky, 2018, and references therein). Satellites

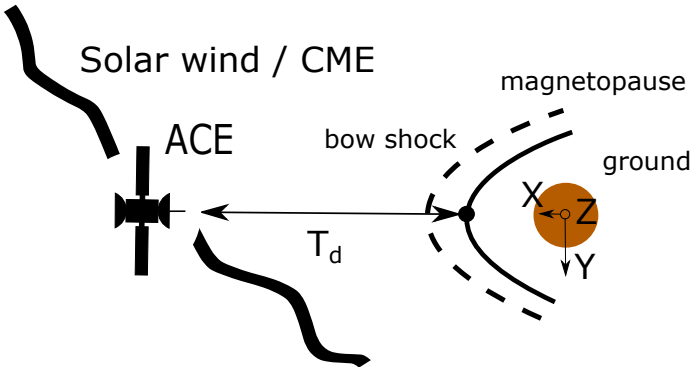

Fig. 3. Schematic of the three-part division of the solar wind propagation delay T_D from L1 (ACE) to Earth. Namely, 1. delay between L1 and bow shock, 2. between bow shock and magnetopause, and 3. between magnetopause and ground, after (Mailyan et al., 2008).

outside of Earth's magnetosphere are able to directly probe the IMF and identify IMF orientation turnings as time stamps for solar wind delay measurements (e.g. Ridley, 2000). However, IMF turnings do not always result in signatures in the magnetosphere which can be used for precise timings of the solar wind propagation delay. Another method is to use interplanetary shock fronts from CME's and CIRs to identify time stamps at solar wind monitor and detector satellites (e.g. Mailyan et al., 2008, and references therein). These shocks are big structures in the solar wind and are unlikely to miss Earth when identified at L1. This study is based on the propagation of interplanetary shocks from ACE to Earth which are visible as sudden commencements at ground based magnetometers (described in Sec. 2).

There are a number of techniques to model the solar wind propagation delay on a physical basis. These techniques can be set into two groups. Firstly, the flat propagation delay which is based on the assumption that the solar wind speed in X-direction is of superior importance for the propagation delay and neglects the other directions. Secondly, a more sophisticated way to derive the time delay is to use the full three-dimensional space instead. Here, the vector of solar wind speed, the normal vector of an interplanetary shock front, the position vector of ACE and a target are taken into account. This method has been termed "vector delay" in the following, as it uses the vector representation of the solar wind propagation delay.

This study introduces a new method based on machine learning and compares this method to the above mentioned physical models of solar wind propagation. In the following, all three methods will be introduced in more detail.

3.1. Flat delay

The simplest way to derive the SW propagation delay, from an L1 spacecraft to the Earth's magnetosphere, is to consider the X-direction only. This approach is called 'flat delay' (e.g. Mailyan et al., 2008) or was termed 'ballistic propagation' in earlier studies. We have adopted the term flat

delay in this study. The assumption that the solar wind speed is dominated by its X-component is the basis of this approach:

$$\Delta t_{flat} = \frac{X_{ACE} - X_T}{v_x}. \tag{2}$$

Here, v_x is the solar wind speed in X-direction, X_{ACE} is the position of ACE along the Earth-Sun line and X_T is the target location. In this study we have used a fixed value for the target location just upstream of Earth, i.e., set X_T to 15 Earth radii (R_E).

The flat delay method has the advantage that it is available as long as there is solar wind speed data from ACE. Its disadvantage is the lack of information on any directionality of the solar wind as well as the interplanetary magnetic field. In addition, the location of ACE around L1 is not fully taken into account.

3.2. Vector delay

A more sophisticated approach to model the SW propagation delay uses all available information from ACE, i.e. full solar wind speed and magnetic field vector. Also, the position of ACE in all three direction is taken into account. In the following, this method will be shortly named vector delay. Its derivation is carried out on the basis of the presence of shocks in the interplanetary medium:

$$\Delta t_{vec} = \frac{(\boldsymbol{r_{ACE}} - \boldsymbol{r_T}) \cdot \boldsymbol{n}}{\boldsymbol{v_{SW}} \cdot \boldsymbol{n}}. \tag{3}$$

Here, $\boldsymbol{r_{ACE}}$ and $\boldsymbol{r_T}$ are the position vector of ACE and the target location. $\boldsymbol{v_{SW}}$ is the three dimensional solar wind vector and $\boldsymbol{n}$ is the normal vector of the interplanetary shock wave heading to Earth. In this study we use a fixed point of the target location and set $\boldsymbol{r_T}$ to (15,0,0) R_E.

The key task of this approach is to determine the normal vector $\boldsymbol{n}$ from ACE measurements. There is a number of techniques available to extract $\boldsymbol{n}$ from solar wind speed and magnetic field measurements. One method is based on the coplanarity assumption, which assumes that the interplanetary shock plane is spanned by two vectors that depend on the magnetic field vector upstream and downstream the interplanetary shock front (e.g. Colburn and Sonett, 1966). A more sophisticated method applies a variance analysis to define $\boldsymbol{n}$ from the minimum variance of the magnetic field (Sonnerup and Cahill, 1967), maximum variance of the electric field or applying a combination of both (Weimer et al., 2003; Weimer and King, 2008). Other methods even solve the full Rankine-Hugoniot problem of the discontinuity to determine the normal vector (Viñas and Scudder, 1986). A good collection and more detail on these methods can be found within the book of Paschmann and Daly (1998).

In this study, we apply the cross product method to derive $\boldsymbol{n}$ (Schwartz, 1998). In the following we will shortly recapitulate the underlying derivation. The coplanarity assumptions allows the definition of $\boldsymbol{n}$ from the following cross products. Magnetic coplanarity (subscript M) yields the following formula for the normal vector:

$$\boldsymbol{n_M} = \pm \frac{(\boldsymbol{B}_d \times \boldsymbol{B}_u) \times \Delta \boldsymbol{B}}{|(\boldsymbol{B}_d \times \boldsymbol{B}_u) \times \Delta \boldsymbol{B}|} \tag{4}$$

$\boldsymbol{B}$ and $\boldsymbol{V}$ are the three dimensional vectors of the magnetic field and solar wind speed measured at ACE. Vectors with subscript d denote downstream conditions and vectors with subscript u denote upstream conditions. The Δ sign indicates the difference between downstream and upstream conditions.

The three following equation rely on the coplanarity of $\boldsymbol{n}$ with a mix of magnetic and solar wind vectors (subscript MX1-3):

$$\boldsymbol{n}_{MX1} = \pm \frac{(\boldsymbol{B}_u \times \Delta \boldsymbol{V}) \times \Delta \boldsymbol{B}}{|(\boldsymbol{B}_u \times \Delta \boldsymbol{V}) \times \Delta \boldsymbol{B}|} \tag{5}$$

$$\boldsymbol{n}_{MX2} = \pm \frac{(\boldsymbol{B}_d \times \Delta \boldsymbol{V}) \times \Delta \boldsymbol{B}}{|(\boldsymbol{B}_d \times \Delta \boldsymbol{V}) \times \Delta \boldsymbol{B}|} \tag{6}$$

$$\boldsymbol{n}_{MX3} = \pm \frac{(\Delta \boldsymbol{B} \times \Delta \boldsymbol{V}) \times \Delta \boldsymbol{B}}{|(\Delta \boldsymbol{B} \times \Delta \boldsymbol{V}) \times \Delta \boldsymbol{B}|} \tag{7}$$

Also the difference between downstream and upstream solar wind speed can be used to derive the normal vector of the interplanetary shock front.

$$\boldsymbol{n}_V = \pm \frac{\boldsymbol{V}_d - \boldsymbol{V}_u}{|\boldsymbol{V}_d - \boldsymbol{V}_u|} \tag{8}$$

Upstream and downstream conditions of $\boldsymbol{B}$ and $\boldsymbol{V}$ have been deduced from averaging measurements 5 min before (upstream) and after (downstream) the shock. For further analysis all five cross product methods (Eq. 4-8) have been evaluated and the mean $\boldsymbol{n}$ has been applied to the derivation of the vector delay (Eq. 3).

The advantage of the vector delay in comparison to the flat delay is its higher accuracy (e.g. Mailyan et al., 2008). The major disadvantage of the vector delay is the requirement of a discontinuity (CME or CIR) within the solar wind to derive $\boldsymbol{n}$. This requirement prevents a timely evaluation of the vector delay and makes its application to a realtime service nearly impossible.

3.3. Machine learning delay

The aim of this new machine learning approach for SW propagation delay modeling is to combine the advantages of the flat and vector delay methods. Specifically, the all-time applicability of the flat delay and the higher accuracy of the vector delay. The all-time applicability is achieved by the nature of the used database. The database only consists of a single ACE datapoint downstream for each interplanetary shock and does not include data from the timeseries several minutes before or after the shock. As a consequence the trained ML model does not know about the presence of a shock front and can be used with continuous data as well. A higher accuracy is expected for the ML approach because the whole position vector of ACE and the Solar wind vector, as similarly applied to the vector delay method, are used for training of the model.

The choice of a machine learning model is often an arbitrary one, mostly dependent upon the computational cost for the specific problem, and in principle many ML algorithms could be applied to the same problem. In this paper we choose to investigate the application of three different ML models in predicting the solar wind propagation delay, namely Random Forest Regression (RF)

(Breiman, 2001), Gradient Boosting (GB) (Friedman, 2001) and Linear Regression (linReg; represented as ordinary least square regression in Pedregosa et al. (2011)).

RF and GB algorithms generate ensembles (forests) of decision trees to make predictions. The main difference between RF and GB model is the characteristics and evaluation of the decision trees in order to produce an output. The RF model builds independent decision trees and produces its result on the basis of an equally weighted average over all trees, a method called bootstrap aggregation in statistics. The GB algorithm improves the performance of individual trees based on a recursive learning procedure, known as boosting. The main reasoning for this choice of models was to firstly enable direct comparison between the RF and GB models, to quantify if the use of an ensemble-based ML model make a significant improvement to the overall performance. Additionally, the linReg model was included as a simple benchmark for comparison with all other models. Both decision tree algorithms exhibit a high degree of versatility and interpretability with regard to the underlying problem, while also demonstrating good overall performance in general (e.g. Biau and Scornet, 2016; Zhang and Haghani, 2015, and references therein). Training and testing of the ML algorithms have been carried out using the Scikit-learn Python package (Pedregosa et al., 2011).

This study uses these machine learning algorithms in their regression representation. Therefore, the machine learning SW propagation delay can be described as the output from a function with the feature vector $\boldsymbol{x}$:

$$\Delta t_{ML} = f_{\mathcal{D}}(\boldsymbol{x}). \tag{9}$$

The notation $f_{\mathcal{D}}$ describes a machine learning algorithm trained on the data set $\mathcal{D}$. The feature vector $\boldsymbol{x}$ contains six features that includes each component of the position vector of ACE (r_x, r_y, r_z) and the measured solar wind speed vector (v_x, v_y, v_z). The data set contains overall 380 samples and is described in Sect. 2. Each sample represents an interplanetary shock measured at ACE and detected as sudden commencement in the magnetosphere. The samples contain the above mentioned feature vector $\boldsymbol{x}$ and the measured solar propagation delay which is the target variable (Y). To avoid biases between different ML models, $\boldsymbol{x}$ is standardized before training.

Hyperparameter optimization

Most machine learning algorithms contain parameters which control their general behavior, the so called hyperparameters. In case of decision tree algorithms, these hyperparameters define the characteristics of the decision trees generated and the number of trees in the forests.

For that purpose Bayesian optimization based on the Gaussian process is often applied (e.g. Swersky et al., 2013, and references therein). This study also follows this paradigm by using the scikit-optimize (Head et al., 2020) python package. For a hyperparameter optimization the database is split into training, testing, and validation set. In this case, the validation set contains 10 % of the database. The remaining 90 % is used for the Bayesian optimization using an internal 5-fold cross-validation. The Bayesian optimization tries to minimize the models root mean square error (RMSE) with respect to the measured SW propagation delay. This is done by consecutively changing the underlying hyper parameters and finding the best set of hyper parameters in this process.

Table 1 shows the hyperparameters that have been taken into account during the optimization of the decision tree models. This table shows also the default parameters used in SciKit-learn. The random forest heavily relies on the number of trees that are generated, as this algorithm reduces

Table 1. Default and Bayesian optimized parameters for the random forest and gradient boosting algorithms, learning rate is not a hyperparameter of the random forest algorithm, minimum impurity decrease was also optimized but did not show a change from 0.0, variable max depth of the default random forest is defined as the expansion of tree until endpoints contain less than [min_samples_split] samples.

Algorithm default/optimized	# Trees	max features per tree	min samples split	min samples leaf	max depth	learning rate
Random Forest	100/800	6/3	2/2	1/1	var./11	N.A.
Gradient Boost	100/490	6/3	2/20	1/8	3/5	0.1/0.02

its output variance by averaging over many random trees. The gradient boosting methods does not require so many trees but the learning rate is important here as it governs the recursive improvement of individual trees. Other hyperparameters define how the branches of the decision trees are generated, i.e. min samples split, min samples leave, max features per tree and max depth of the decision trees. For a closer description of these hyperparameters we refer to the SciKit-learn documentation (Pedregosa et al., 2011, scikit-learn 0.22.1).

The next step is to investigate the performance of the optimized ML algorithm compared to the default algorithm. For that purpose we performed this time a k-fold cross-validation, using 10 folds. Here, the full dataset is split into 10 parts where each segment is iteratively used as the test set with the other remaining segments used to train the model. To prevent biases due to training data being ordered in time, the database has been randomly shuffled first. The segmentation is then done in a stratified way, so that each fold contains a training data distribution(shock cases) that represents the full range of measured solar wind propagation delays. The RMSE with respect to the measured SW delay acts as a performance metric. Figure 4 shows all ten folds in a set of histograms. The best algorithm to predict the SW delay is the optimized gradient boosting method with a mean RMSE of 4.5 min, closely followed by the optimized random forest with 4.7 min. That is an improvement of 8 % and 5 % for the optimized gradient boosting and random forest algorithm, respectively, compared to their default counterparts. The GB method benefits more from hyperparameter optimization than the random forest, however the RF in its default version performs slightly better than the default GB. Applying a cross validation shows that some SW propagation delay cases of the database are more difficult to predict than others. This behavior is also independent of whether the ML algorithms are used with optimized or default hyperparameters. In the following we will compare the optimized ML results to the flat and vector delay method.

4. Results

4.1. Comparison of ML and physical delay models

At first we investigate the performance of all delay model is investigated for the example case shown in Fig. 1. Table 2 summarizes the delay model outputs for this case and its actual SW propagation delay measurement of. The ML models were trained with the whole database excluding only the case on 2000/07/19 and used the feature vector (248 R_E, 13 R_E, 18 R_E, -556 km/s, -76 km/s, 91 km/s) for its prediction. The RF and GB model as well as the vector method predict the SW propagation

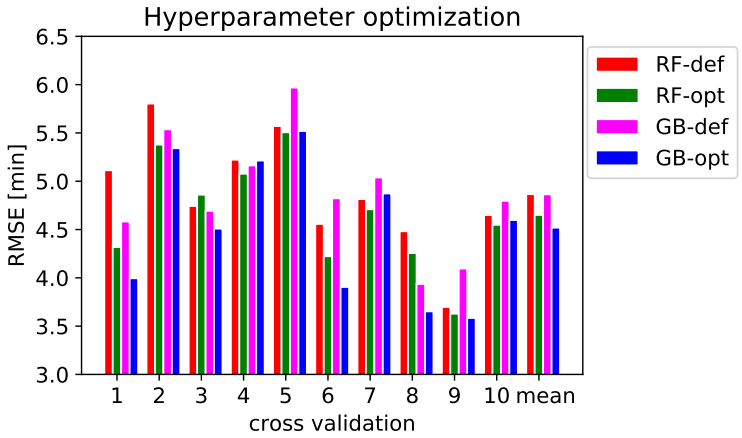

Fig. 4. Histogram of RMSE for the 10 fold cross validation of default (red) and optimized (green) random forest as well as the default (magenta) and optimized (blue) gradient boosting algorithm, the fold number is indicated on the x-axis together with the mean of all folds.

Table 2. Solar wind propagation delay predictions using all five methods for the example CME on 2000/07/19 (see Fig. 1) with feature vector (248 R_E, 13 R_E, 18 R_E, -556 km/s, -76 km/s, 91 km/s). The two digits after the decimal point are insignificant, as the measurement error is in the order of two minutes.

delay model	RFreg	GBreg	vector	flat	linReg	measured
SW delay [min]	37.(85)	38.(45)	39.(05)	44.(23)	43.(12)	38.(41)

delay for this specific CME with less than one minute deviation from the measurement. The linear regression and flat delay method overestimate the delay by 5 respectively 6 min.

For an statistical assessment of the ML performance we use the cross-validation approach used in the hyperparameter optimization (see Fig. 4). Stratified K-fold cross validation is a robust method to investigate the performance of a ML algorithm. This approach prevents positive bias when interpreting the statistical nature of the ML results compared to other methods. The comparison contains three ML algorithms, i.e. random forest, gradient boost and linear regression, and the flat and vector method to model the SW propagation delay. Taking simple linear regression into account allows us to investigate if the more sophisticated ML algorithms achieve greater performance when compared with a very simplistic model.

In order to achieve a reasonable comparison of the trained ML algorithms to the flat and vector methods to model the solar wind propagation delay, we use the same test sets to derive RMSEs for all methods. Figure 5 contains the results of the 10-fold cross-validation for all five methods to model the SW propagation delay. Here, the metrics (RMSE, MAE and mean error) serve as an accuracy measure for each method's capability to predict the solar wind propagation delay. The gradient boosting and random forest results are the same as in Fig. 4 for the optimized algorithms. The linear regression model has been trained and tested with the same cross-validation folds. The performance of the physical models is based on the test sets only.

The comparison using RMSE metric (Fig. 5 a) reveals that the decision tree models (RF and GB) boosting perform better than both the LinReg and Fphysical models. The ten-fold cross-validation shows that RF and GB perform almost the same with a variation between 5.5 and 3.5 min and a mean of only 4.7 and 4.5 min, respectively. The best physical method to model the SW delay is the vector method with a mean RMSE of 5.3 min. However, its range is also higher, ranging from 6.5 min down to 4.1 min. Linear regression performs with a mean RMSE of 6.1 min in the fourth place. The flat method shows the worst result among all studied algorithms and can model the solar wind delay with a mean RMSE of only 7.3 min.

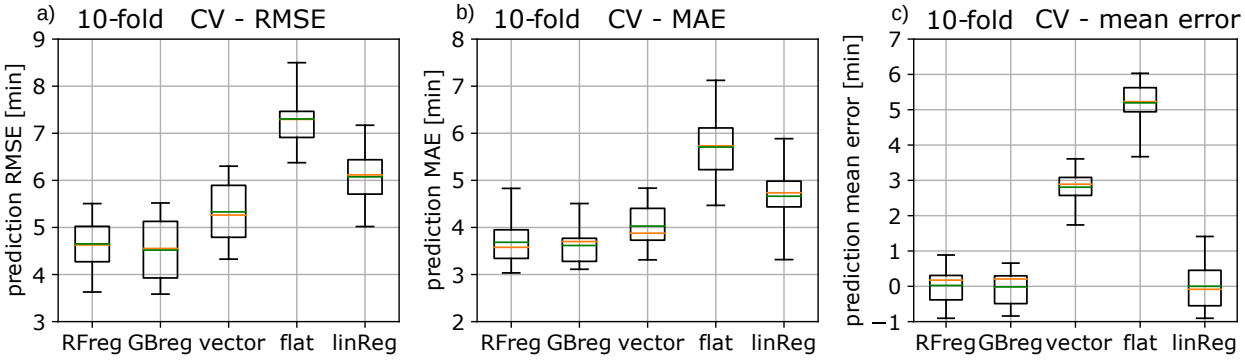

Fig. 5. Comparison of model performance based on a 10-fold cross validation (CV) between random forest (RFreg), gradient boost (GBreg), vector delay, flat delay, and basic linear regression (linReg) shown as box plot for the performance metrics a) RMSE , b) mean absolute error (MAE) and c) mean error . Each box contains 10 folds from the cross validation, its mean is shown in green, the median in orange, the box edge show the 25/75 percentile and the whiskers indicate the full range.

The same cross validation, but using the mean absolute error as a metric (Fig. 5 b), show a similar ranking of the different approaches to predict the SW propagation delay. The mean error metric (Fig. 5 b) shows a different behaviour. All ML models show mean errors around 0 minutes, which is expected from their statistical nature. Only the physical models show a bias when considering the mean error. The vector delay overestimates the SW delay by 2.8 min while the flat delay overestimates by even 5.2 min.

Figure 6 shows a comparison of the flat method and fully trained GB algorithm predictions based on ACE level 2 data from the first 100 days in 2019. Interplanetary shocks from 2019 onward are not part of the training data set, the chosen period is therefore completely unknown to the GB model. This comparison points at the ultimate future application for machine learning based SW propagation delay predictions, i.e. a realtime operational setting. It has to be noted that the analysis in Fig.6 is only qualitative as no ground truth of the solar wind propagation delay is available to the authors. A rigorous validation of this kind of continuous application is subject to a future study.

During the majority of the time period, both predictions are qualitatively similar and vary between 30 and 70 min. However, when the solar wind speed in Y and Z-direction is non-zero the trained GB model predicts propagation delays several minutes shorter than the flat method. Between 15. March and 01. April there is a time period of solar wind speed as low as -250 km/s in the X-direction. While the flat delay predicts propagation delays of up to 90 min, the GB model output stays around 77 minutes. This behavior can be explained by the lack of training data for these very

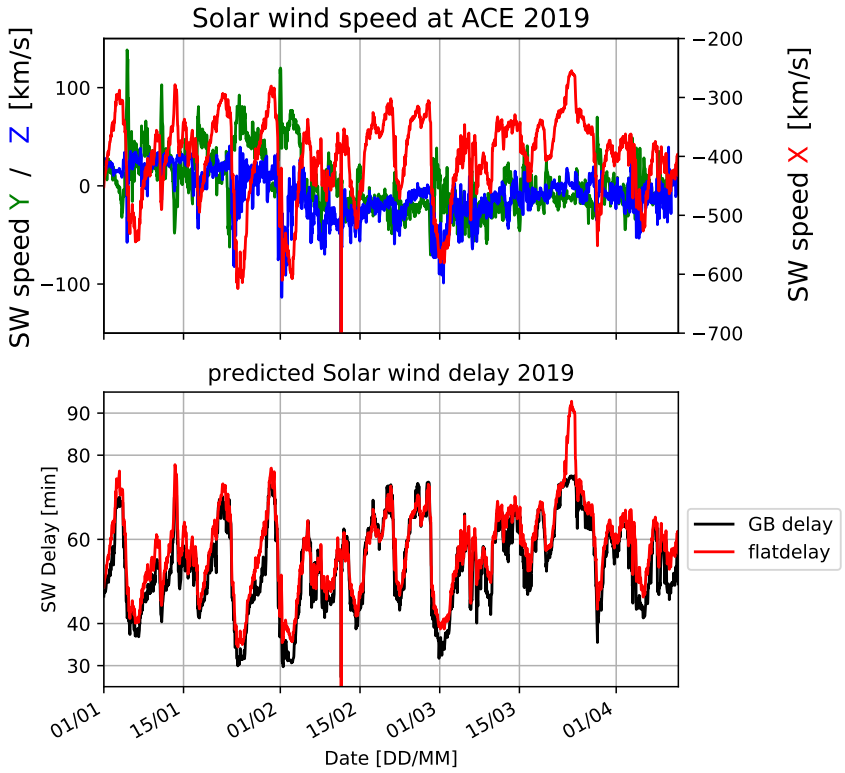

Fig. 6. Comparison of SW propagation delay prediction for the beginning of 2019 using the flat delay method and a fully trained GB model. Continuous ACE level 2 hourly data is used for input.

low values of solar wind speed. In the context of this work, there is no evidence that any heliospheric shock with solar wind speed at this low level generates a detectable sudden commencement in the magnetosphere.

The further analysis concentrates on the GB machine learning model only. The RF as well as the linear regression are discarded from that analysis.

4.2. Performance Validation

For the purpose of implementing these machine-learning models in predicting the solar wind propagation delay, it is important to consider performance validation of the entire dataset. In the field of machine learning it is standard practice to choose a train/test ratio of 80/20 (i.e., 80 % of the data is training and 20 % is testing) or 90/10 when verifying the performance of models. The choice of these ratios is typically subjective and provides only a single-valued estimate of each model's performance. An analysis of the dependence of the performance metric (RMSE) on the selection of training/testing set ratios is presented in Fig. 7. As the flat and vector method are simply analytical models based on physical assumptions, they do not rely on statistical training, evidenced by the near-constant RMSE value for the variable test sets.

As expected, the performance of the GB model depends significantly on training set size and tends to converge toward a stable RMSE when the training set contains at least 40 samples. Already with this small amount of training data, the GB model can predict the SW propagation delay with

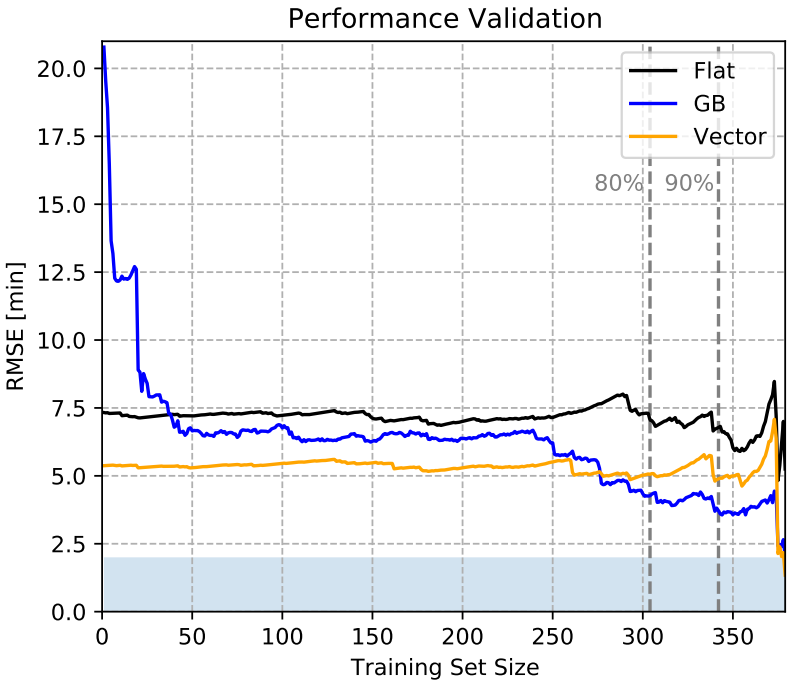

Fig. 7. Comparison of model performance (RMSE) with varying train/test set ratios for the gradient boosting (blue) model. Physical models, vector delay (orange) and flat delay (black), are evaluated on the same test set size which is equal to 380 - training set size. The shaded region corresponds to the estimated error in the delay measurement.

a lower RMSE than the simple flat method. Otherwise, relying on these small training set sizes the GB model cannot outperform the vector delay that achieves lower values of RMSE, i.e., better performance.

However, the performance of the GB model begins to increase again when training set size reaches ≈ 250 cases of the total data set. Leading to ultimately better performance of the GB model than either the flat or vector methods when using more than 270 cases for training. It is also interesting to note that as the testing set size decreases to $< 10\,\%$ of total data set size, all model RMSE values increase, i.e, their performance decreases. This behaviour can be explained by considering the case of low-number statistics, the testing set size is not sufficiently large enough and therefore the RMSE values have increasingly high variance. Hence, performance values in this range are deemed statistically unreliable.

Using the standard 80/20 split case, the flat, vector and GB models achieve RMSE values of 7.1, 5.1 and 4.3 min respectively. In the case of a 90/10 split, the flat, vector and GB models achieve RMSE values of 6.8, 4.9 and 3.7 min, respectively. In both cases, the GB model out-performs both the vector and flat methods, a result that was previously reflected in the k-fold cross validation analysis (see Fig.5).

4.3. Explaining the gradient boosting results

To improve the understanding of the physical mechanisms, information from the trained gradient boosting algorithm is extracted. To start with, Fig. 8 shows the correlation matrix of the used features based on the underlying database (cf. 2) and the target value (T_D). There are two combinations of enhanced correlation among the features itself. Firstly, the position of ACE in X and Z-direction have a correlation index of 0.44. This correlation originates from the nature of the ACE' orbit around L1. Overall, the dataset is only slightly correlated and all features are expected to contribute to the prediction based on the machine learning model. In addition, the solar wind speed in X-direction is strongly correlated ($c = 0.85$) to the measured solar wind propagation delay, that is expected from the assumption in the flat delay approach. The other features are hardly correlated to the target variable. The identified correlations has to be taken into account in the further explanation of the trained GB model.

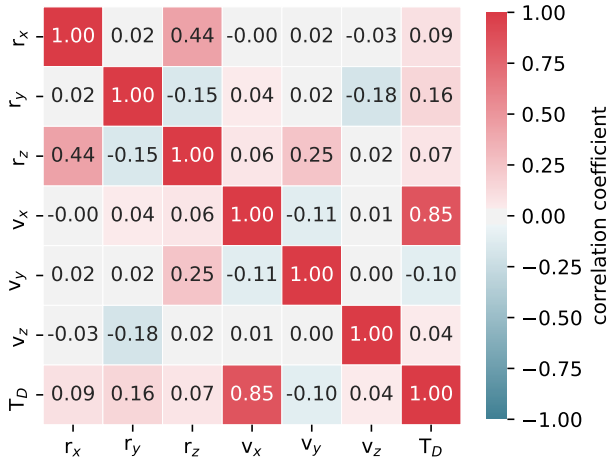

Fig. 8. Correlation matrix of the machine learning features (position of ACE (r_x, r_y, r_z), solar wind speed (v_x, v_y, v_z)) and the measured solar wind delay T_D of the database.

One way of explaining trained machine learning algorithms is to derive the so-called feature importance (FI). FI is usually derived to identify the subset of features which has the biggest impact on ML model accuracy and robustness. Selecting only the most valuable and relevant features also decreases the time needed to train ML models.

However, there are many different interpretations of how FI can be retrieved from machine learning algorithms (e.g. Hapfelmeier and Ulm, 2013, and references therein). This study performs drop column FI as this method is able to identify unambiguously the feature importance from random forests (Strobl et al., 2007). Drop-column FI determines the change in performance when a feature (column) is left out (dropped) of the feature set to train the GB model when compared to a fully trained model. As a performance metric the RMSE is used here again.

Drop-column FI values can be positive and also negative. Positive values indicate that leaving out a certain feature increases the RMSE of the ML model. Features showing a negative FI indicate that leaving out this feature reduces the RMSE of machine learning model, i.e. the performance

increases. The drop column feature importance (DC_{FI}) of feature x for a trained ML algorithm can be represented as follows:

$$DC_{FI}(x) = RMSE(x \notin F) - RMSE(x \in F). \tag{10}$$

Here, $RMSE(x \notin F)$ is the RMSE obtained from a trained random forest leaving feature x out of the used feature set F. $RMSE(x \in F)$ is the RMSE of the fully trained random forest. Both RMSE's are evaluated from the same test dataset.

Figure 9 shows the results of drop-column FI determination. To identify the statistical behaviour of the 6 features of the random forest model, a 10-fold cross-validation is performed for the drop column FI.

Each box in Fig. 9 contains the mean importance as well as its variability for each feature. By far the highest increase in RMSE occurs when the solar wind speed in X-direction v_x is left out of the feature set, FI values range between 3 and 6 min with a mean of 4.6 min and the median at 4.5 min. All other features show mean importances below one minute, some folds within the cross validation show even negative values. The solar wind speed components in Y and Z-direction(v_y, v_z) show smaller feature importance ($\approx$ 30 sec), however all train/test folds show positive values. The FI of the position of ACE is even lower. The FI of the ACE position X and Y-component have positive mean values around 10-20 s. For the slightly correlated feature r_z we find an importance close to zero, i.e. r_z does not contribute to the performance of the trained random forest.

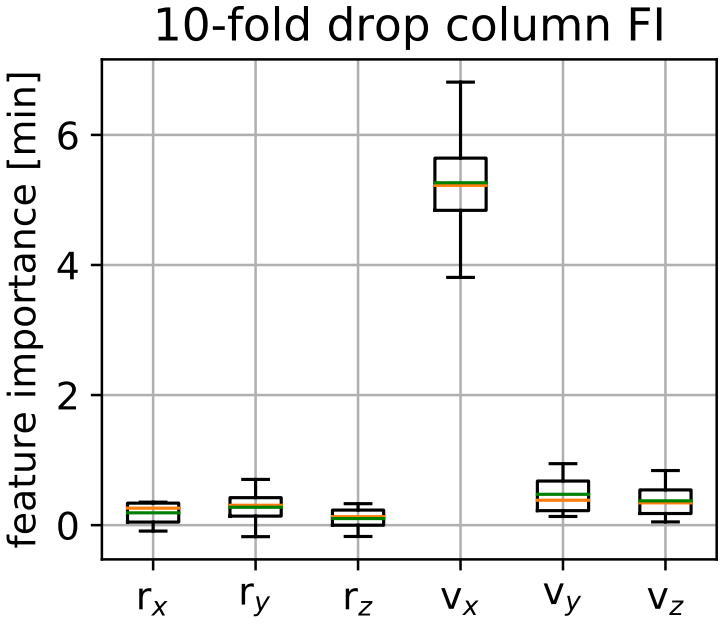

Fig. 9. Drop column feature importance of the GB model using 10-fold cross validation, whispers show the full range, the box show the 25 %/75 % percentiles, the green line indicates the mean, the orange line indicates the median.

Drop-column FI only gives a general view of the trained GB model, but the underlying functioning of the algorithm remain unresolved. In order to get a glimpse into the GB itself, Shapley

values can open a view into its depth. Shapley (1953) proposed a measure to identify the bonus due to cooperation within a cooperative game. The surplus that each player contributes to the outcome of the game is called the Shapley value. The principle of a cooperative game can also be applied to the GB regression of this study. Here, the ML features resemble Shapley's cooperative players. The python package SHAP provides functionalities to derive Shapley values from trained ML algorithms (Lundberg and Lee, 2017) and was also used for the random forest analysis in this study.

A fully trained GB model, i.e, all features and samples have been used for training, is the basis of the Shapley value analysis. Shapley values can be calculated for each sample of the database (Sect. 2) In this study the Shapley values represent the changes of the predicted solar wind propagation delay with respect to a mean model output. A negative Shapley values indicate that the model output for this individual sample results in a shorter solar wind delay compared to the mean model output. This is the case for positive values, but the individual model output is higher compared to the mean model solar wind delay. Table 3 shows the mean values of all features obtained from the database. For example, the mean speed of the interplanetary shocks detected at ACE is -469 km/s in X-direction. The trained random forest predicts a solar wind propagation delay of 47 min when these mean feature values are used for input to the random forest. The following scatter plots contain Shapley values for all 380 individual samples in the database.

Table 3. Mean model input values (not standardized) and mean propagation delay from the fully trained Random Forest model.

$\overline{v_X}$ [km/s]	$\overline{v_Y}$ [km/s]	$\overline{v_Z}$ [km/s]	$\overline{r_X}$ [RE]	$\overline{r_Y}$ [RE]	$\overline{r_Z}$ [RE]	$\overline{T_{delay}}$ [min]
-469	-14	0.6	233	0.89	0.29	47

Figure 10 panel a) shows the Shapley values for solar wind speed in X-direction, v_x, which has the biggest feature importance within the GB model. The Shapley value reaches +/-20 min for cases with v_x = -300 km/s and -900 km/s respectively. The relationship between v_x and the Shapley values follows the assumption of constant solar wind speed and can be described by the function $t(v) = r/v - t_0$. Here, t_0 can be identified as the mean solar wind delay and r as the distance between ACE and the magnetopause. The blue line in Figure 10 a) indicates the best fit to the distribution of Shapley values. The fit yields values for t_0= 43 min and r= 198 R_E which are close to the actual mean values in Tab. 3.

The Shapley value for solar wind speed in Y (v_y) and Z-direction (v_z) look completely different, see Fig. 10 b), c). An important difference to v_x is that $v_y$and v_z vary between positive and negative values. In both cases the relationship between shapley value and v_y/v_z seem to have a quadratic form. The Shapley value are negative for cases with high absolute solar wind speeds in Y and Z-direction, they can reach down to -6 min. Slightly positive Shapley values (< 2 min) group around v_y/v_z being close to zero. The shapley values for v_y are shifted to negative solar wind speeds, i.e. close to the mean value of v_y of -14 km/s. For v_z, the parabola is closely centered around zero solar wind speed in Z-direction, as is the mean solar wind speed in that direction.

Shapley values for the position of ACE (r_x,r_y,r_z) are shown in Figure 10 d) ,e) ,f). As ACE orbits around L1, it moves from 215 to 250 R_E in r_x, within +/- 50 R_E in r_y, and within +/-25 R_E in r_z (see also Fig. 2). The relationship between r_x and the resulting Shapley values is linear and shows

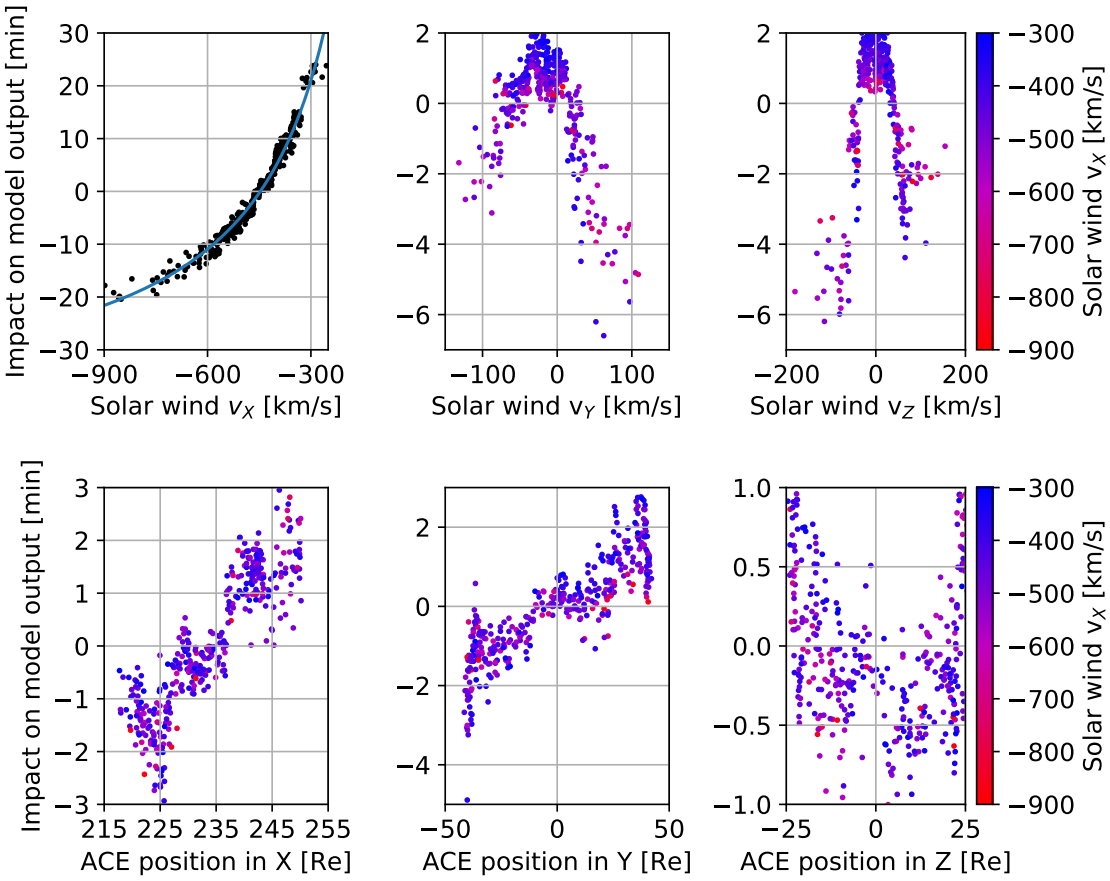

Fig. 10. Shapley values for all six ML features, a) solar wind speed in X-direction, b) solar wind speed Y-direction, c) solar wind speed Z-direction, d) ACE position in X-direction, e) ACE position in Y-direction, f) ACE position in Z-direction, color-coded is the solar wind speed in X-direction.

up to +/- 1 min impact on model output for higher/smaller distances from Earth. A similar linear relationship is exhibited by r_y and its Shapley values, here negative r_y values correspond to negative Shapley values of up to -2 min and vice versa. However, this linearity is affected by the solar wind speed in X-direction, v_x. In the case of high solar wind speed, the Shapley values for r_y are below ± 2 min. When v_x is rather small, the model output is more affected by r_y, especially when ACE is far from the Sun-Earh line. For r_z no linear relationship to the derived Shapley can be seen. As already seen by the drop-column FI, also the Shapley value for r_z are not much higher than ± 0.5 min. When the ACE satellite is off the Sun-Earth line, the Shapley value reaches higher absolute values but the Shapley value can be both negative and positive.

5. Discussion

The results described in Sec. 4 show the possibilities of machine learning for the modeling of solar wind propagation delays. Here, we discuss and interpret the scientific impact of these results.

This study shows the possibility to time the solar wind propagation delay between a solar monitor at L1 and the magnetosphere. We used the magnetosphere's ability to act as a detector of interplan-

etary shocks for the timing of the solar wind propagation delay. Precise timing is possible with the help of magnetometers not only onboard of satellites but also on Earth's surface.

It has to be noted here that setting the target location rigidly to (15,0,0) R_E introduces an additional error to the physical models. The time delays used in this study are however closely related to the interaction of the solar wind with the magnetosphere near the magnetopause. The location of the magnetopause varies between 6 and 15 R_E (e.g. Sibeck et al., 1991) and setting a fixed target location introduces an error in the order of one minute for the physical models. A bias has been identified in the mean error of the vector (2.8 min) and flat delay method (5.2 min) (see Fig. 5 c)), which is influenced by the target location as well. Setting the target location to 10 R_E would have further increased that bias.

However, the comparison between physics based modeling of SW propagation delay and trained ML algorithms remains fair because the ML algorithms do not have specific information about the location of the magnetopause either. Cash et al. (2016) even use 30 Earth radii for their target to determine the solar wind delay, to account for property changes of the solar wind as it approaches the bow shock.

The cross validation shows a better performance of the trained gradient boosting model compared to the vector delay method. However, it has to be noted that the representation of the vector delay in this study might not be optimal. There are many parameters, e.g., the underlying technique to evaluate the normal vector or the number of data points used to define upstream and downstream conditions of the interplanetary shock, that govern the performance of the vector delay. A very detailed optimization of these parameters might further increase the performance of the vector delay but this is not the scope of this manuscript.

The good performance of the GB model is a favorable result by its own, but its biggest advantage is the versatility of the ML approach. While, the vector delay method depends on a rigorous analysis of solar wind data during shock events, the ML approach only needs values for its six features at a single point in time to output a solar wind delay. The training database consists of interplanetary shock events detected at ACE only, however the trained model can predict the arrival time of solar wind features of any kind at any given time. That allows for its application in a near real-time warning service for users in the space industries.

As the database used in this study relies on the timing of CMEs and CIRs only, a full generalization of the ML approach to times without interplanetary shocks remains to be investigated. Cash et al. (2016) examined the generalization of the vector delay based on the minimum variance technique to a real time application, this study should be used as a blue print for the investigation on the ML approach. It has to be noted that the ML approach has deficiencies (see Fig. 6) when the input data is outside of the parameterspace used for training. This finding is in line with Smith et al. (2020), who showed that ML based classification of sudden commencements can face misclassifications when applied outside of the trained parameter space.

A thorough analysis of the trained GB model improves the confidence in its solar wind propagation delay output. That analysis includes an operational validation analysis with various train / test splits, a drop column feature importance analysis and a Shapley value analysis.

In the case of the operational validation analysis (Fig. 7), GB performs better than both vector and flat methods when choosing an 80/20 or 90/10 split of the training and validation sets. Even though the vector method performs relatively well in its prediction, with sufficient training (more

than 200 cases of the database), the GB model achieves greater performance overall and has the potential for greater improvement if more training data instances were available.

The importance of the six features of the dataset has been analysed based on a drop-column feature analysis. All features contribute to the performance of the solar wind propagation delay. The most important feature is the solar wind speed in X-direction, what is also expected from this problem.

Furthermore, the Shapley value analysis of the trained GB algorithm also gives additional confidence in the prediction of the solar wind propagation between L1 solar monitors and Earths magnetosphere. From Fig. 10 a,b,c) one can summarize that high solar wind speeds in any direction lead to a shorter propagation delay output from the GB model. From Fig. 10 d) it is obvious that a shorter distance between satellite and Earth corresponds to a shorter propagation delay output.

Slightly different is the case of the Y-component of ACE position around L1. From Fig. 10 e) it is obvious that the Y-component of the ACE position can increase, as well as decrease the modelled propagation delay. The decrease of propagation delay for negative values of r_y and an increase for positive values can be accounted to the Parker spiral nature of the solar wind. Especially low speed cases show this effect. High solar wind speed cases are less affected by the Parker spiral effect on the solar wind propagation delay. A similar finding was shown by Mailyan et al. (2008, Fig.4) in their flat delay analysis, which shows a linear dependence of their flat delay error on the difference between the Y-component of the position of ACE and Cluster position. The mean solar wind speed in the Y-direction of all cases in the database is also negative (-14 km/s cf. Tab. 3). This shift and its representation in the Shapley values of Fig. 10 b) can be interpreted as an effect of the Parker spiral nature of the solar wind as well.

The DSCOVR satellite has been in L1 orbit since 2015 and will be the only solar monitor after the decommissioning of ACE and Wind. As the orbits of DSCOVR and ACE are very similar, we expect that our trained algorithm is also capable of predicting the solar wind propagation delay from DSCOVR data with similar accuracy. However, this has to be validated and is not within the scope of this manuscript.

Real-time predictions of SW propagation delay needs NOAA's real-time solar wind (RTSW) data. The results shown Fig. 6 can be seen as the first successful demonstration of the continuous application of the ML approach. However, the provided data is of different nature compared to the Level 2 ACE data used in this study. The ACE RTSW data provides only bulk solar wind speed, proton density and three components of the magnetic field strength. In addition to that, RTSW data suffers from higher noise levels and additional data gaps. These problems impact the results of SW propagation delay predictions for ML as well as physical models. A future study will construct a new database based on RTSW data and investigate if a trained ML algorithms on that RTSW data can outperform the simple flat delay method. Additionally, the ML model could also benefit from information on the location of the magnetopause prior to a CME impact.

A future study can also investigate the role of the magnetic field before and after a interplanetary shock event to improve the predictions. Solar wind information upstream and downstream of a shock can be used as additional ML features. By doing so, it can be investigated if giving additional information on the shock's normal vector could further improve the machine learning performance and provide a fairer comparison to the vector delay methods. Furthermore, using the solar wind pressure may also be a feature to be included into the ML approach for better predictions, since it would help to indicate the position of the magnetopause.

The newly introduced ML approach to predict solar wind propagation delays from L1 data can be put into the group of velocity-based approaches like the flat and vector delay. A key drawback of simple velocity-based delay methods is that SW properties propagated to the target position can arrive out of order. There are schemes that already exist that address the problem of an unphysical propagation structure (e.g., OMNI, Weimer and King (2008)). Hydrodynamic models on the other hand, which generate continuous data outputs based on the physical evolution of the plasma structure at L1, do not suffer from this drawback.

6. Conclusions

This work shows the possibility to model the solar wind propagation delay between L1 and Earth on the basis of machine learning. A database has been generated on the basis of ACE data and ground based magnetometer data which served as the training set for the random forest ML algorithm. This database contains 380 measurements of the solar wind propagation delay from the detection of interplanetary shocks at ACE and their signature as sudden commencement in Earth's magnetosphere.

Random forest, gradient boosting and linear regression have been applied to identify a suitable model for the SW propagation delay. Here, the gradient boosting algorithm performs best (RMSE = 4.5 min), closely followed by the random forest and with larger margin, the linear regression. We also performed a hyperparameter optimization and found a slight improvement of 5-8 % to the default ML algorithms.

We performed a comparison of the ML model to physical models to derive the solar wind propagation delay, i.e., flat delay and vector delay. The trained GB algorithm performs significantly better than the flat propagation delay model, i.e., the RMSE for the flat delay is more than 2 min larger than for the GB approach. The comparison showed that the vector delay method performs slightly worse compared to the trained GB model with an RMSE of 5.3 min. An application of the GB model to continuous solar wind data revealed that the predictions follow the flat delay as long as the solar wind speed in y and z is close to zero. The GB predictions give a shorter propagation time than the flat delay when that it is not the case. In addition, this analysis revealed that the GB output is closely related to the underlying training data. When operated outside the trained parameter space, e.g. when the solar wind is below 300 km/s, the GB model gives out unrealistic propagation delays.

The analysis of the trained GB algorithm was performed on the basis of a performance validation, feature importances and Shapley values. The performance validation revealed that the GB model needs to be trained with at least 200 cases of the database in order to perform at par with the vector delay method. In addition, the feature importance and Shapley value analysis enhanced the confidence in the trained GB algorithms and its predictions. The solar wind speed in X-direction was identified as the most important feature in the feature set. The Shapley value analysis revealed the internal relationship between the features and also indicated that the trained GB model follow basic physical principles like an empirical model.

The trained GB algorithm is suited to be run for post-event analysis or with near real-time data. The trained algorithm only needs input for solar wind speed vector and ACE's position vector to predict the solar wind propagation delay reliably.

Acknowledgements. The authors thank Dymtro Vasylyev and Leonie Pick for valuable discussions on the topics of machine learning and Earth's magnetic field.

This paper uses data from the Heliospheric Shock Database (ipshocks.fi), generated and maintained at the University of Helsinki. We acknowledge the provision of ACE data by the ACE science Center at Caltech (www.srl.caltech.edu/ACE). The magnetometer data used in this work was made available by the INTERMAGNET consortium (intermagnet.org). We also acknowledge the SuperMAG service (supermag.jhuapl.edu) for magnetometer data visualisation. This research made use of HelioPy, a community-developed Python package for space physics (Stansby et al., 2019).

The authors would like to thank two anonymous referees and Enrico Camporeale for their constructive comments which have helped to improve the manuscript.

References

Araki, T., 1977. Global structure of geomagnetic sudden commencements. *Planetary and Space Science*, **25**(4), 373 – 384. 10.1016/0032-0633(77)90053-8. 1

Baker, D. N., W. K. Peterson, S. Eriksson, X. Li, J. B. Blake, et al., 2002. Timing of magnetic reconnection initiation during a global magnetospheric substorm onset. *Geophysical Research Letters*, **29**(24), 43–1–43–4. 10.1029/2002GL015539. 1

Baumann, C., and A. E. McCloskey, 2020. Measurements of the solar wind propagation delay for L1 to Earth based on ACE and ground-based magnetometer data. 10.5281/zenodo.4300253. 2

Biau, G., and E. Scornet, 2016. A random forest guided tour. *TEST*, **25**(2), 197–227. 10.1007/s11749-016-0481-7. 3.3

Borovsky, J. E., 2018. The spatial structure of the oncoming solar wind at Earth and the shortcomings of a solar-wind monitor at L1. *Journal of Atmospheric and Solar-Terrestrial Physics*, **177**, 2–11. Dynamics of the Sun-Earth System: Recent Observations and Predictions, 10.1016/j.jastp.2017.03.014. 3

Breiman, L., 2001. Random Forests. *Machine Learning*, **45**(1), 5–32. 10.1023/a:1010933404324. 3.3

Cameron, T., and B. Jackel, 2016. Quantitative evaluation of solar wind time-shifting methods. *Space Weather*, **14**(11), 973–981. 10.1002/2016sw001451. 1

Cameron, T. G., and B. Jackel, 2019. Using a Numerical MHD Model to Improve Solar Wind Time Shifting. *Space Weather*, **17**(5), 662–671. 10.1029/2019sw002175. 1

Camporeale, E., 2019. The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather*, **17**(8), 1166–1207. 10.1029/2018SW002061. 1

Cash, M. D., S. W. Hicks, D. A. Biesecker, A. A. Reinard, C. A. de Koning, and D. R. Weimer, 2016. Validation of an operational product to determine L1 to Earth propagation time delays. *Space Weather*, **14**(2), 93–112. 10.1002/2015sw001321. 1, 5

Colburn, D. S., and C. P. Sonett, 1966. Discontinuities in the solar wind. *Space Science Reviews*, **5**, 439–506. 10.1007/BF00240575. 3.2

Curto, J. J., T. Araki, and L. F. Alberca, 2007. Evolution of the concept of Sudden Storm Commencements and their operative identification. *Earth, Planets and Space*, **59**(11), i–xii. 10.1186/bf03352059. 1, 2

Engebretson, M. J., D. L. Murr, W. J. Hughes, H. Lühr, T. Moretto, et al., 1999. A multipoint determination of the propagation velocity of a sudden commencement across the polar ionosphere. *Journal of Geophysical Research: Space Physics*, **104**(A10), 22,433–22,451. 10.1029/1999ja900237. 2

Friedman, J. H., 2001. Greedy function appoximation: A gradient boosting machine. *The Annals of Statistics*, **29**(5), 1189–1232. 10.1214/aos/1013203451. 3.3

Gjerloev, J. W., 2012. The SuperMAG data processing technique. *Journal of Geophysical Research: Space Physics*, **117**(A9). 10.1029/2012JA017683. 2

Gosling, J. T., J. R. Asbridge, S. J. Bame, A. J. Hundhausen, and I. B. Strong, 1967. Discontinuities in the solar wind associated with sudden geomagnetic impulses and storm commencements. *Journal of Geophysical Research*, **72**(13), 3357–3363. 10.1029/jz072i013p03357. 1

Haaland, S., C. Munteanu, and B. Mailyan, 2010. Solar wind propagation delay: Comment on "Minimum variance analysis-based propagation of the solar wind observations: Application to real-time global magnetohydrodynamic simulations" by A. Pulkkinen and L. Raststätter. *Space Weather*, **8**(6), n/a–n/a. 10.1029/2009sw000542. 1

Haiducek, J. D., D. T. Welling, N. Y. Ganushkina, S. K. Morley, and D. S. Ozturk, 2017. SWMF Global Magnetosphere Simulations of January 2005: Geomagnetic Indices and Cross-Polar Cap Potential. *Space Weather*, **15**(12), 1567–1587. 10.1002/2017SW001695. 1

Hapfelmeier, A., and K. Ulm, 2013. A new variable selection approach using Random Forests. *Computational Statistics & Data Analysis*, **60**, 50 – 69. 10.1016/j.csda.2012.09.020. 4.3

Head, T., M. Kumar, H. Nahrstaedt, G. Louppe, and I. Shcherbatyi, 2020. scikit-optimize/scikit-optimize. 10.5281/zenodo.4014775. 3.3

Horbury, T. S., D. Burgess, M. Fränz, and C. J. Owen, 2001. Three spacecraft observations of solar wind discontinuities. *Geophysical Research Letters*, **28**(4), 677–680. 10.1029/2000gl000121. 3

Jian, L., C. T. Russell, J. G. Luhmann, and R. M. Skoug, 2006. Properties of Interplanetary Coronal Mass Ejections at One AU During 1995 – 2004. *Solar Physics*, **239**(1-2), 393–436. 10.1007/s11207-006-0133-2. 2

Kömle, N. I., H. I. M. Lichtenegger, and H. O. Rucker. The Sun and the Heliosphere in Three Dimensions, chap. Propagation of Solar Wind Features: A Model Comparison Using Voyager Data. Astrophysics and Space Science Library, 1986. 10.1007/978-94-009-4612-5_26. 1

Liu, J., Y. Ye, C. Shen, Y. Wang, and R. Erdélyi, 2018. A New Tool for CME Arrival Time Prediction using Machine Learning Algorithms: CAT-PUMA. *The Astrophysical Journal*, **855**(2), 109. 10.3847/1538-4357/aaae69. 1

Love, J. J., and A. Chulliat, 2013. An International Network of Magnetic Observatories. *Eos, Transactions American Geophysical Union*, **94**(42), 373–374. 10.1002/2013EO420001. 2

Lundberg, S. M., and S.-I. Lee, 2017. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Advances in Neural Information Processing Systems 30, 4765–4774. Curran Associates, Inc. 4.3

Mailyan, B., C. Munteanu, and S. Haaland, 2008. What is the best method to calculate the solar wind propagation delay? *Annales Geophysicae*, **26**(8), 2383–2394. 10.5194/angeo-26-2383-2008. 1, 3, 3, 3.1, 3.2, 5

McComas, D. J., S. J. Bame, P. Barker, W. C. Feldman, J. L. Phillipsa, P. Riley, and J. W. Griffee, 1998. Solar Wind Electron Proton Alpha Monitor (SWEPAM) for the Advanced Composition Explorer. *Space Science Reviews*, **86**, 563–612. 10.1023/A:1005040232597. 2

Oliveira, D. M., and J. Raeder, 2015. Impact angle control of interplanetary shock geoeffectiveness: A statistical study. *Journal of Geophysical Research: Space Physics*, **120**(6), 4313–4323. 10.1002/2015ja021147. 2

Paschmann, G., and P. W. Daly, 1998. Analysis Methods for Multi-Spacecraft Data. ISSI Scientific Reports Series SR-001, ESA/ISSI, Vol. 1. ISBN 1608-280X, 1998. *ISSI Scientific Reports Series*, **1**. 3.2

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830. 3.3, 3.3

Pulkkinen, A., and L. Rastätter, 2009. Minimum variance analysis-based propagation of the solar wind observations: Application to real-time global magnetohydrodynamic simulations. *Space Weather*, **7**(12), n/a–n/a. 10.1029/2009sw000468. 1

Ridley, A. J., 2000. Estimations of the uncertainty in timing the relationship between magnetospheric and solar wind processes. *Journal of Atmospheric and Solar-Terrestrial Physics*, **62**(9), 757–771. 10.1016/s1364-6826(00)00057-2. 1, 3

Schwartz, S. J. Analysis Methods for Multi-Spacecraft Data, chap. Shock and Discontinuity Normals, MachNumbers, and Related Parameters, 249–305. ISSI Scientific Report, 1998. 3.2

Segarra, A., M. Nosé, J. J. Curto, and T. Araki, 2015. Multipoint observation of the response of the magnetosphere and ionosphere related to the sudden impulse event on 19 November 2007. *Journal of Space Weather and Space Climate*, **5**, A13. 10.1051/swsc/2015016. 2

Shapley, L. S., 1953. A value for n-person games. *Contributions to the Theory of Games*, **2**(28), 307–317. 4.3

Sibeck, D. G., R. E. Lopez, and E. C. Roelof, 1991. Solar wind control of the magnetopause shape, location, and motion. *Journal of Geophysical Research*, **96**(A4), 5489. 10.1029/90ja02464. 5

Smith, A. W., I. J. Rae, C. Forsyth, D. M. Oliveira, M. P. Freeman, and D. R. Jackson, 2020. Probabilistic Forecasts of Storm Sudden Commencements From Interplanetary Shocks Using Machine Learning. *Space Weather*, **18**(11). 10.1029/2020sw002603. 1, 5

Smith, C. W., J. L'Heureux, N. F. Ness, M. H. Acuña, L. F. Burlaga, and J. Scheifele. The Ace Magnetic Fields Experiment, 613–632. Springer Netherlands, Dordrecht, 1998. ISBN 978-94-011-4762-0. $10.1007/978-94-011-4762-0_21$. 2

Sonnerup, B. U. O., and L. J. Cahill, 1967. Magnetopause structure and attitude from Explorer 12 observations. *Journal of Geophysical Research*, **72**(1), 171. 10.1029/jz072i001p00171. 3.2

Stansby, D., Y. Rai, J. Broll, S. Shaw, and Aditya, 2019. HelioPy: Heliospheric and planetary physics library. 1903.017. 6

Stone, E., A. Frandsen, R. Mewaldt, E. Christian, D. Margolies, J. Ormes, and F. Snow, 1998. The Advanced Composition Explorer. *Space Science Reviews*, **86**(1/4), 1–22. 10.1023/a:1005082526237. 1

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn, 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25. 4.3

Swersky, K., J. Snoek, and R. P. Adams, 2013. Multi-Task Bayesian Optimization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., Advances in Neural Information Processing Systems 26, 2004–2012. Curran Associates, Inc. 3.3

Viñas, A. F., and J. D. Scudder, 1986. Fast and optimal solution to the "Rankine-Hugoniot problem". *Journal of Geophysical Research: Space Physics*, **91**(A1), 39–58. 10.1029/JA091iA01p00039. 3.2

Weimer, D. R., and J. H. King, 2008. Improved calculations of interplanetary magnetic field phase front angles and propagation time delays. *Journal of Geophysical Research: Space Physics*, **113**(A1), n/a–n/a. 10.1029/2007ja012452. 1, 3.2, 5

Weimer, D. R., D. M. Ober, N. C. Maynard, M. R. Collier, D. J. McComas, N. F. Ness, C. W. Smith, and J. Watermann, 2003. Predicting interplanetary magnetic field (IMF) propagation delay times using the minimum variance technique. *Journal of Geophysical Research: Space Physics*, **108**(A1). 10.1029/2002JA009405. 3, 3.2

Wu, C., C. D. Fry, D. Berdichevsky, M. Dryer, Z. Smith, and T. Detman, 2005. Predicting the Arrival Time of Shock Passages at Earth. *Sol. Phys.*, **227**, 371–386. 10.1007/s11207-005-1213-4. 1

Yang, Y., F. Shen, Z. Yang, and X. Feng, 2018. Prediction of Solar Wind Speed at 1 AU Using an Artificial Neural Network. *Space Weather*, **16**(9), 1227–1244. 10.1029/2018sw001955. 1

Zhang, Y., and A. Haghani, 2015. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, **58**, 308–324. 10.1016/j.trc.2015.02.019. 3.3

Zhelavskaya, I. S., R. Vasile, Y. Y. Shprits, C. Stolle, and J. Matzka, 2019. Systematic Analysis of Machine Learning and Feature Selection Techniques for Prediction of the Kp Index. *Space Weather*, **17**(10), 1461–1486. 10.1029/2019sw002271. 1