Solar Physics (2024) 299:95
https://doi.org/10.1007/s11207-024-02339-2

RESEARCH

# Comparative Analysis of Various Machine-Learning Models for Solar-Wind Propagation-Delay Estimation

**Hemapriya Raju**[1] · **Saurabh Das**[1]

Received: 6 December 2023 / Accepted: 20 June 2024 / Published online: 4 July 2024


## Abstract

Geomagnetic storms resulting from solar disturbances impact telecommunication and satellite systems. Satellites are positioned at Lagrange point L1 to monitor these disturbances and give warning 30 min to 1 h ahead. As propagation delay from L1 to Earth depends on various factors, estimating the delay using the assumption of ballistic propagation can result in greater uncertainty. In this study, we aim to reduce the uncertainty in the propagation delay by using machine-learning (ML) models. Solar-wind velocity components ($V_x$, $V_y$, $V_z$), the position of Advanced Composition Explorer (ACE) at all three coordinates ($r_x$, $r_y$, $r_z$), and the Earth's dipole tilt angle at the time of the disturbances are taken as input parameters. The target is the time taken by the disturbances to reach from L1 to the magnetosphere. The study involves a comparison of eight ML models that are trained across three different speed ranges of solar-wind disturbances. For low and very high-speed solar wind, the vector-delay method fares better than the flat-plane propagation method and ML models. Ridge regression performs consistently better at all three speed ranges in ML models. For high-speed solar wind, boosting models perform well with an error of around 3.8 min better than the vector-delay model. Studying the best-performing models through variable-importance measures, the velocity component $V_x$ is identified as the most important feature for the estimation and aligns well with the flat-plane propagation method. Additionally, for slow solar-wind disturbances, the position of ACE is seen as the second most important feature in ridge regression, while high-speed disturbances emphasize the importance of other vector components of solar-wind speed over the ACE position. This work improves our understanding of the propagation delay of different solar-wind speed and showcases the potential of ML in space weather prediction.

**Keywords** Coronal mass ejections · Interplanetary · Solar wind · Disturbances

✉ H. Raju
phd1901121008@iiti.ac.in; hemapriya.rceg@gmail.com

S. Das
saurabh.das@iiti.ac.in

[1] Department of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology Indore, Madhyapradesh 453552, India

## 1. Introduction

The solar wind is a continuous stream of plasma emanating from the Sun. Sudden eruptions such as coronal mass ejections (CMEs) or corotating interaction regions (CIRs) can interact with the Earth's magnetosphere. Subsequently, the shock or the disturbances compress the Earth's magnetosphere, due to sudden changes in solar-wind dynamic pressure. Ground-based magnetometer stations can detect these variations as sudden changes in the horizontal component of the Earth's magnetic field, and this phenomenon is termed as sudden commencements (Nishida, Cahill, and Laurence 1964; Burlaga and Ogilvie 1969; Joselyn and Tsurutani 1990; Fujita 2019). Sudden commencements may or may not be followed by geomagnetic storms, depending on the southward-oriented interplanetary magnetic field (IMF) $B_z$ component (Gosling et al. 1967; Tsurutani and Baker 1979).

These storms pose risks to satellite systems, communication networks, and power grids, thus emphasizing the importance of estimating the solar-wind parameters such as velocity, magnetic-field strength, and orientation of the IMF $B_z$ component at the Earth's bowshock. Continuous monitoring of these disturbances by a single satellite at a fixed location near the Earth's bowshock can be challenging. Thus, satellites positioned at the relatively stable Lagrangian L1 orbit, 1.5 million km ahead of Earth, help us to provide uninterrupted monitoring of solar-wind disturbances and also have the advantage of producing sufficient time ahead for the warnings (Crooker et al. 1982).

The large-scale solar-wind structures identified at L1 have a high probability that they will hit Earth, thus giving us advanced time for the preparation. However, the timing with which it reaches the Earth's atmosphere varies from around 30 min to 120 min depending on their speed of arrival and various other factors. Once the interplanetary shock is identified by Advanced Composition Explorer (ACE), the onset time of the propagation delay at L1 is defined as when the solar-wind speed reaches its peak value. Thus, the propagation delay is considered from the time taken for the disturbances from L1 to reach the magnetosphere, which is observed as the sudden commencements using the ground-based magnetometer stations. The accurate arrival time of these disturbances is needed, as the solar-wind parameters are provided as input for magnetosphere coupling models, ionospheric and thermospheric variations, and substorm studies (Ridley et al. 1998; Lavraud et al. 2006). Timely data is essential for safeguarding orbiting satellites and preventing ground-induced currents (GICs) in power grids.

The well-defined methods for estimating the propagation delay, still have significant uncertainty in predicting their arrival. A simplified flat-plane delay model considers calculating the propagation delay with the solar-wind velocity, assuming that the disturbances travel normal to the flat plane in the line connecting L1 to the Earth. However, the accuracy of the flat-plane propagation model deviates from the actual observations arriving at the Earth, that can go beyond 60 min which can be observed from the long tail from the Gaussian distribution of the propagation delay (Collier et al. 1998). Thus, considering only distance along the Sun–Earth line ignores the following factors:

i) The solar-wind properties are monitored by the satellite ACE, orbiting around Lagrangian L1 point around 1.5 million km from Earth (McComas et al. 1998). However, the spacecraft orbits around L1 and varies about $\pm 40$ $R_E$ in the $Y$ direction, adding additional uncertainty about the arrival of the solar-wind disturbances, which needs to be studied in detail (Ridley 2000; Milan et al. 2022).
ii) The solar-wind disturbances follow the parker spiral. Thus, the IMF plane fronts can have tilted orientations that can vary up to 45 deg from the flat plane of the L1 to the Earth line, increasing uncertainty in the propagation delay (Kelly et al. 1986).

iii) Also, the interaction of the solar-wind plasma, when high-speed solar-wind speed compresses the slow solar-wind varies the arrival time of these disturbances.

Several studies have been carried out to improve the prediction of the propagation delay by considering the tilt of the IMF plane front (Russell, Siscoe, and Smith 1980). Thus, estimating the normal phase front along the parker spiral is essential. For scenarios involving multiple satellites, constructing a normal phase front is achievable through geometric means (Blanchard and Bankston 2002). In the case of single-point observations, the normal phase front is calculated using the minimum variance analysis (MVA) method by choosing a normal vector, where the boundary magnetic-field structure variations are minimized (Lepping and Behannon 1980; Sonnerup and Scheible 1998). The normal vector can be estimated either through the crossproduct of the average magnetic-field vectors upstream and downstream of the discontinuities (Horbury et al. 2001) or by computing the eigenvector associated with minimum eigenvalue that will be oriented perpendicular to the IMF mean field vector, and considered as the plane of the IMF phase front (Weimer et al. 2003; Weimer and King 2008). Despite these advancements, accuracy diminishes with increased spacecraft distance when moving away from L1 to the Earth line (Collier et al. 1998; Ridley et al. 1998; Weimer et al. 2002; Milan et al. 2022). Further, these models assume that the solar wind with constant speed, neglecting interactions between high-speed and slow-speed solar winds. The nature of the disturbances, the evolution of the solar wind from L1 to Earth, and the shocks formed are well studied through magnetohydrodynamic (MHD) simulations, considering the propagation of disturbances in 1-dimension through the MVA method from L1 to the Earth (Pulkkinen and Rastätter 2009; Cameron and Jackel 2019). Munteanu et al. (2013) performed wavelet denoising to eliminate small-scale fluctuations of the boundary normal estimations of the magnetic-field structures and the disturbances are propagated through the MVAB method to improve the accuracy of existing propagation delay-estimation methods. In a comparative analysis conducted by Mailyan, Munteanu, and Haaland (2008), Cash et al. (2016), and Cameron and Jackel (2016) on various propagation delay-estimation methods, it was observed that techniques such as MVA and others prove effective in mitigating uncertainty in propagation. However, the flat-plane propagation method seems more feasible to implement for real-time applications.

Baumann and McCloskey (2021) proposed a machine-learning (ML) model to calculate the propagation delay of the shock events of the solar disturbances with input parameters as the position of ACE in all three coordinates ($r_x$, $r_y$, $r_z$) and all three components of solar-wind velocity vectors ($V_x$, $V_y$, $V_z$). The Gradient Boost regressor performs better than the flat-plane propagation delay and vector-delay method with root mean square error (RMSE) of around 4.5 min. O'Brien et al. (2023) introduced a probabilistic approach for regression for estimation of the disturbances at the magnetosphere by including the past history of the solar wind from L1 as input parameter. The model performs better than the MVA method and can provide physically meaningful uncertainties.

In this work, following the work of Baumann and McCloskey (2021), we predict the propagation delay between the shock structures observed at L1 and disturbances observed by ground-based magnetometers as sudden commencements using eight ML models. The input of the models is the position of ACE in three coordinates, and solar-wind velocity vectors of all three components. In addition, we have added the tilt angle of the Earth's magnetic axis as the input features. The dipole tilt angle affects the position and strength of magnetic reconnection (Eggington et al. 2020). The magnetopause topology is highly sensitive to dipole tilt angle, as seen by MHD simulations (Maynard et al. 2001). Thus, the dipole tilt angle of the Earth can impact both the reconnection component and shift the first point of contact of the solar wind with the magnetopause nose to northward or southward

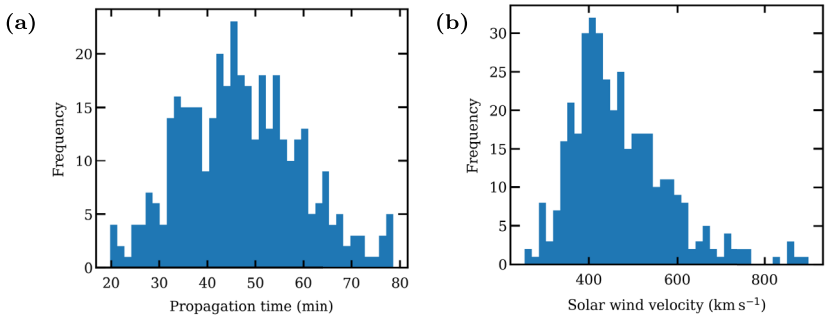


**Figure 1** (a) Distribution of time taken for solar-wind disturbances to reach from L1 to Earth. (b) Distribution of solar-wind velocity ($V_x$) as observed at L1.

(Liu et al. 2012; Lu et al. 2013; Zhu et al. 2015). In this work we have carried out the following:

- We have added the additional variable, dipole tilt angle, presuming that the duration of the solar-wind interaction with the magnetosphere may vary depending on the tilt of Earth's magnetic dipole. We try to use all these seven parameters as input to our ML models.
- Based on solar-wind velocity ($V_x$), we have divided the data into three velocity bins: less than 400 km s$^{-1}$, between 400 and 600 km s$^{-1}$, and more than 600 km s$^{-1}$. We compare the performance of the ML models on different solar-wind speeds, with conventional methods such as flat-propagation delay and vector-based delay methods.
- We study the important features that contribute to the different solar-wind speed using interpretable machine learning with permutation-based variable-importance measure.

This article is divided into the following sections. Section 2 explains the nature of the data and its source, model overview, and the training process. Section 3 discusses the results of the ML models and the interpretation of the relevant features from the training. Section 4 discusses the overall results, and the scope of future work.

## 2. Data Source and Methodology

The data source includes interplanetary shocks to estimate the propagation delay and spans the period from 1998 to 2018, encompassing a total of 380 events (Baumann and McCloskey 2021). In this work, instead of continuous solar-wind parameters, only shock events caused by CME disturbances and CIRs are considered. The solar-wind velocity of all three vector components ($V_x$, $V_y$, $V_z$) is obtained from ACE observed at the downstream of the shock structures. The position of ACE in all three coordinates ($r_x$, $r_y$, $r_z$) and the speed are given as input parameters to the machine-learning models. Further, the Earth's dipole tilt affects the way the solar wind and magnetopause interact. Thus, an additional variable of Earth's dipole tilt angle corresponding to the shock-event date is added as the input feature using the dipole Python package (Karl Laundal et al. 2020). As mentioned in the introduction, the propagation delay is observed as the time difference between the shock arrival at L1 and the onset of sudden commencements as observed by ground-based magnetometers. The distribution of the propagation time is shown in Figure 1(a). It can be seen that the disturbances take around 20 min, in the case of the high-speed solar wind, and up to 80 min in the case of slow solar wind to reach the Earth from L1. The average time is around 40 to 50 min

to reach the Earth. The distribution of the solar-wind speed for these disturbances is shown in Figure 1(b), where the majority of solar-wind speed falls in the range of 400 km $s^{-1}$ to 600 km $s^{-1}$.

### 2.1. Model Overview

We feed the input parameters and the propagation time into the eight various ML models and compare their performance with base models such as the simple flat-plane propagation model and the vector-based delay model. The flat-plane-based propagation-delay method is based on the simple ballistic propagation of L1 to the Earth by considering only the velocity $V_x$ component (Collier et al. 1998). This is given by Equation 1:

$$\text{Propagation delay} = \frac{\text{Distance from L1 to the Earth}}{\text{Solar-wind velocity } (V_x)}. \tag{1}$$

The vector-based delay method considers the positions of ACE in all three coordinates, thereby considering the orientation of the normal vector of the shock. The normal vector is calculated by the crossproduct between the upstream and downstream discontinuities identified (Schwartz 1998; Horbury et al. 2001; Baumann and McCloskey 2021):

$$\text{Propagation delay} = \frac{(\text{Position of ACE at L1} - \text{Earth's location}).\hat{n}}{(V_{sw}).\hat{n}}. \tag{2}$$

The machine-learning models Ridge regression, Random Forest (RF), Decision Tree (DT), adaboost, Support Vector Regression (SVR), XGBoost Regressor (XGBR), Gradient Boost Regressor (GBR), and MultiLayer Perception (MLP) models are trained for the prediction of solar-wind propagation delay.

Ridge regression is a modified linear regression with a regularized method efficient for the data exhibiting multicollinearity (McDonald 2009). Thus, it will be more robust and efficient in generalization. SVR models are regression models that use a hyperplane to fit the data, allowing support vectors at the same time within a certain margin of error (Awad and Khanna 2015). Thus, it helps in the generalization of the unseen data. SVR is also efficient in handling nonlinear relationship between features and the target through kernels. DT tries to discriminate the classification, starting from the root node and descending down to the leaf node by exhibiting certain decisions. Thus, it is robust against data with correlated features but can be biased easily and leads to overfitting. To make it efficient, we can use $n$ decision trees in the form of boot strapping and take the average of the outcomes, thus helping it to become a more generalizable method called random forest (Ho 1995).

Boosting models work on weighted averages in the sequential adjustments instead of averaging the outcomes. Adaboost or adaptive boosting uses several weak learners and weights are adjusted sequentially, thereby trying to reduce misclassification in each iteration. Thus, it is very sensitive to outliers and easily prone to overfitting. GBR is another type of boosting method, which uses weak learners in iterations to update the model parameters by fitting it to the residuals in each step of iterations and thus trying to minimize the loss function (Friedman 2002). This model is robust against outliers. The XGBR model is similar to gradient boosting, but fast and efficient method with regularization parameters and inbuilt optimization methods.

MLP is a kind of feedforward neural network where it contains a few layers of neurons or perceptrons that are fully interconnected to provide output (Popescu et al. 2009). The input neuron is fed with input features, weights, and biases with nonlinear activation and passed

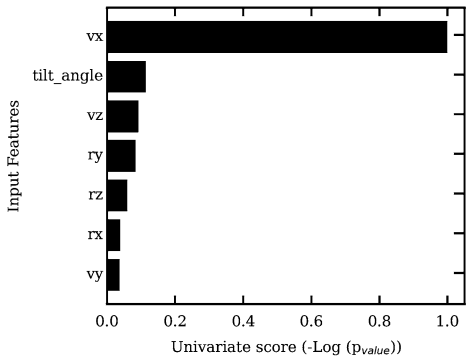


**Figure 2** Feature importance using the Select Kbest method to study the relevant features of the propagation.

through a feedforward propagation method to other neurons till the output node or neuron is reached. The weights and biases are adjusted through the subsequent iterations through the backpropagation method and error is minimized using a loss function.

## 2.2. Feature Importance and Training Methodology

Before training, the individual features contributing to the estimation of propagation delay are studied through feature scores from the *Select Kbest* method, thereby assessing the important features. *Select Kbest* computes the crosscorrelation between the features and the target value. The resultant correlation values are transformed into feature scores and P-values. Subsequently, the values are reordered and sorted with the highest feature relevance at the top. Figure 2 provides a visual representation of this ranking and shows that the solar-wind velocity ($V_x$) component is identified as the most relevant feature, followed by the Earth's dipole tilt angle, which is seen as a significant predictor for the propagation delay. We aim to assess the improvement in prediction performance by employing a combination of these features as input for the ML models.

The dataset taken ($V_x$, $V_y$, $V_z$, $r_x$, $r_y$, $r_z$, tilt angle) are scaled using mean-based standardization separately to each column of input features, as mentioned in Equation 3. The $X_i$ is the individual data point of a respective feature. $\bar{X}$ and $\sigma$ are the mean and standard deviation of the data points for a respective feature, as shown in Equation 3:

$$X_{\mathrm{i(standardized)}} = \frac{X_{\mathrm{i}} - \bar{X}}{\sigma}. \tag{3}$$

Once scaling of the input features is done, we employ a $k$-fold cross-validation method to split the dataset randomly into $k$ number of folds, train on $k$-1 folds, and test the trained model on the remaining $k$th fold for each ML model. This process of training is carried out iteratively till all the $k$ folds are tested. Calculating the mean and standard deviation of test results of $k$ folds gives us the model uncertainty on unseen data. Further, results of vector-based delay and flat-plane propagation delay are also considered for splitting into folds along with train and test set, so as to compare the accuracy with both these methods for those respective test dataset. Training and testing as a whole dataset is carried out for 3-fold cross-validation.

Slow solar wind undergoes many variations before reaching Earth, due to its interactions with high-speed solar-wind disturbances and other factors, thus presenting higher uncertainty. To check the performance of the models for different solar-wind speeds of various

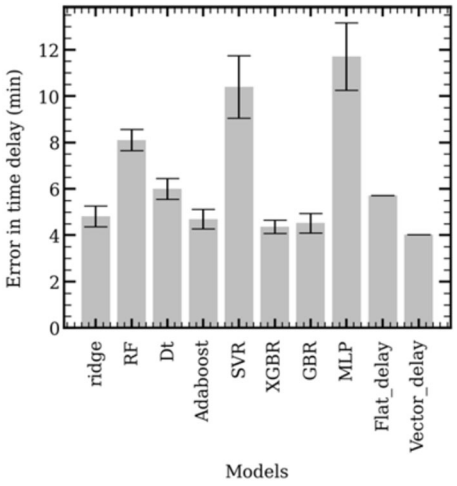


**Figure 3** Propagation-delay error of ML models, flat-delay model, vector-delay models are compared. The absolute values represent the mean, and the error bar represents the standard deviation of the propagation-delay error obtained for the entire dataset tested across 3-fold cross-validation.

solar-wind disturbances, we partition the dataset into three bins based on the solar-wind velocity $V_x$. The first bin comprises data of slow solar-wind speed, with speed below 400 km $s^{-1}$. The second bin has those data where the solar-wind speed lies between 400 km $s^{-1}$ and 600 km $s^{-1}$. The third bin has those disturbances with speed greater than 600 km $s^{-1}$. Among the dataset of 380 data points, 107 data points belong to bin1, 233 data points belong to bin2, and 40 data points belong to bin3. We treat each velocity bin as a separate dataset and train the ML models separately. We employ 2-fold cross-validation for this training. The comparative analysis is done further to study the propagation-delay error of the machine-learning models and base models.

## 3. Results and Discussion

In this section, we discuss the results of the ML models and aim to comprehend the important features of the best-performing models. Figure 3 represents the propagation-delay error of each ML model trained and tested on the entire dataset. The results are compared with the flat-delay model and vector-delay models. The final results are the mean of the propagation-delay error and the error bars represent the standard deviation obtained across the 3-folds that tells us the uncertainty of the ML model for the unseen data. As we can observe, the flat-plane method has a propagational delay error of around 5.5 min. The vector-based propagation method gives a better delay error of around 4 min. Boosting models such as XGBR and GBR perform well only next to vector-based propagation method, but better than the flat-propagation method.

In Figure 4, we analyze the variations in each model's performance with the dataset that was separated based on the velocity ranges. For each bin, the final results are the mean values, and the error bars represent the standard deviation of the propagation-delay error obtained across 2-fold cross-validation. Overall, we see that the propagation-delay error decreases with increasing solar-wind speed. The Ridge regression model performs well with less error than other ML models for slow solar-wind speed, but next to the performance of the vector-based propagation-delay method. For the solar-wind speed that lies between 400 and 600 km $s^{-1}$, Adaboost, XGBR, and GBR models perform better than the base models with the majority of the data points in this speed range. For very high-speed solar wind,

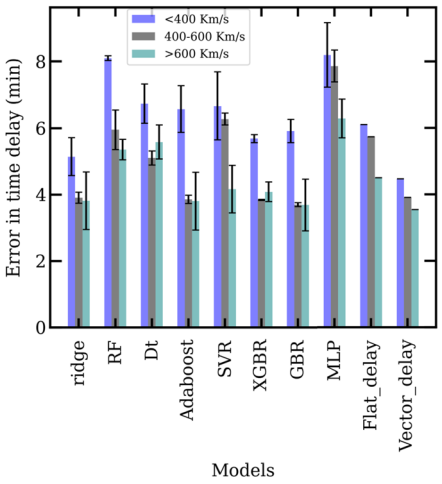


**Figure 4** Propagation-delay error of ML models, flat-delay model, vector-delay models are compared for the dataset that is split based on different speed ranges. The absolute values represent the mean, and the error bar represents the standard deviation of the propagation-delay error obtained across 2-fold cross-validation for the respective speed bins.

**Table 1** Comparison of propagation-delay error (min) of base models with machine-learning models.

| Model | Speed $< 400$ km s$^{-1}$ | Speed of $400-600$ km s$^{-1}$ | Speed $> 600$ km s$^{-1}$ |
|---|---|---|---|
| Flat delay | 6.09 | 5.73 | 4.49 |
| Vector delay | 4.46 | 3.90 | 3.54 |
| Ridge | $5.14 \pm 0.57$ | $3.90 \pm 0.86$ | $3.81 \pm 0.86$ |
| RF | $8.09 \pm 0.07$ | $5.95 \pm 0.59$ | $5.35 \pm 0.30$ |
| DT | $6.74 \pm 0.59$ | $5.10 \pm 0.21$ | $5.58 \pm 0.51$ |
| Adaboost | $6.57 \pm 0.70$ | $3.85 \pm 0.12$ | $3.80 \pm 0.87$ |
| SVR | $6.67 \pm 1.01$ | $6.27 \pm 0.17$ | $4.16 \pm 0.71$ |
| XGB | $5.68 \pm 0.11$ | $3.84 \pm 0.019$ | $4.08 \pm 0.29$ |
| GBR | $5.91 \pm 0.34$ | $3.69 \pm 0.06$ | $3.68 \pm 0.78$ |
| MLP | $8.19 \pm 0.17$ | $7.86 \pm 0.48$ | $6.28 \pm 0.58$ |

Ridge regression, Adaboost, XGBR, and GBR models perform well compared to other ML models and base models. As we had done cross-validation of 2-folds, the uncertainty of the model comes from the results of 2-folds that tells us about the generalization of the model. The detailed results are shown in Table 1.

Since we observe a good variation in the error between the slow solar-wind component and the other higher solar-wind speed disturbances, we study the features that are relevant to this variation using interpretable ML methods. This is done by variable-importance measure, which involves dropping a feature, carrying out the model's prediction and the respective loss is calculated as dropout loss. Then, the variables are sorted in descending order according to their dropout loss. The more loss a variable incurs, the more important the feature is for the model's performance. This can be used to study the underlying reason for the model's performance too. The work uses the dalex Python module to carry out this variable-importance measure (Baniecki et al. 2021).

Here, we have chosen the ML models ridge, adaboost, XGBR, and GBR, which, when compared to flat-plane propagation delay, and vector-based delay models, show relatively better performance than other ML models. The important features are analyzed for these

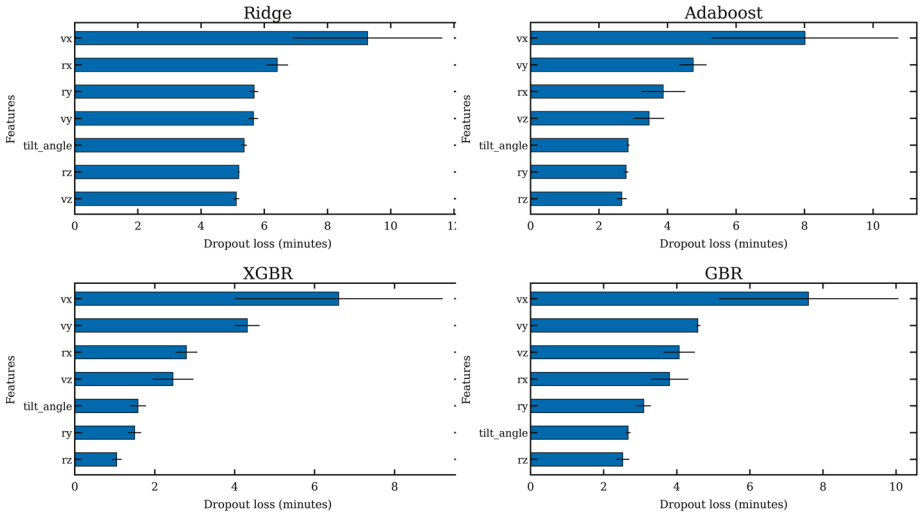


**Figure 5** Feature importance for bin1 comprising data of solar-wind speed less than 400 km $s^{-1}$. The error bar represents the standard deviation of the propagation-delay error obtained across 2-fold cross-validation for the bin1 dataset.

machine-learning models. Figure 5 shows the important features of those ML models trained with a dataset comprising slow solar-wind speed. The mean and standard deviation are obtained across 2-fold cross-validation and represented as the absolute values and the uncertainty of the ML models for the unseen data. The ridge model performs better than other ML models and the flat-delay method with propagation-delay error of around 5.14 min. We observe that $V_x$ leads as the important feature for the prediction, agreeing with the flat-plane propagation method. The next two important features were the position of ACE ($r_x$, $r_y$) in the Ridge regression model. Other ML models show other components of the speed vector that may be the reason for their lower performance. In slow solar-wind speed, the tilt angle also plays an important role in prediction next to the position of ACE, as seen in Figure 5.

Figure 6 shows the performance of the selected ML models for the speed range that lies between 400 km $s^{-1}$ and 600 km $s^{-1}$. In this speed range, ridge gives similar error as the vector-delay method at around 3.9 min as the delay error. Adaboost and XGBR show lesser error of around 3.85 min, whereas GBR shows the least error among all ML models in this speed range of around 3.69 min when compared to the vector-delay method as shown in Figure 4. Analyzing the feature importance, $V_x$ remains as the dominant feature in all models, thus we discuss the remaining features that are next relevant. For bin2, ridge, adaboost, XGBR and the GBR models show $V_z$ as the next important feature that incurs greater loss next to $V_x$. The model GBR that fares better with lesser propagation-delay error, shows that the first 3 important features are the velocity components of the solar-wind followed by the position of ACE. Figure 7 shows the performance of selected models for a speed range greater than 600 km $s^{-1}$. In this speed bin, GBR performs similar to the vector-delay method, while GBR has a delay error around 3.68 min and the vector-delay method has a delay error of 3.54 min. Ridge and adaboost have delay errors of around 3.8 min and XGBR has a delay error of 4.08 min. GBR shows a $V_z$ component incurring greater loss next to $V_x$ as similar to the speed bin2, and followed by the position of ACE ($r_x$). As the dataset for this bin is very small, the results need to be studied for consistency with larger datasets. For high and very high-speed solar wind, other components of solar-wind velocity are observed

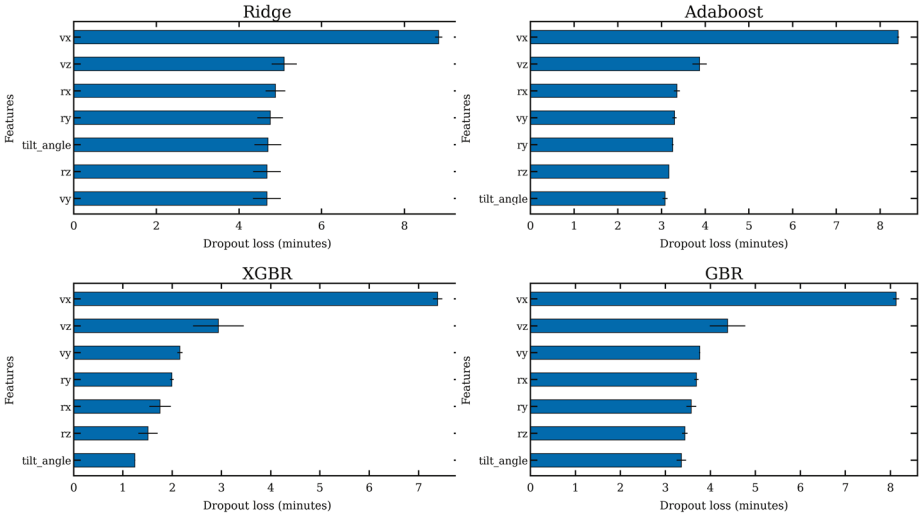


**Figure 6** Feature importance for bin2 with solar-wind speed between 400 km $s^{-1}$ and 600 km $s^{-1}$. The error bar represents the standard deviation of the propagation-delay error obtained across 2-fold cross-validation for the bin2 dataset.

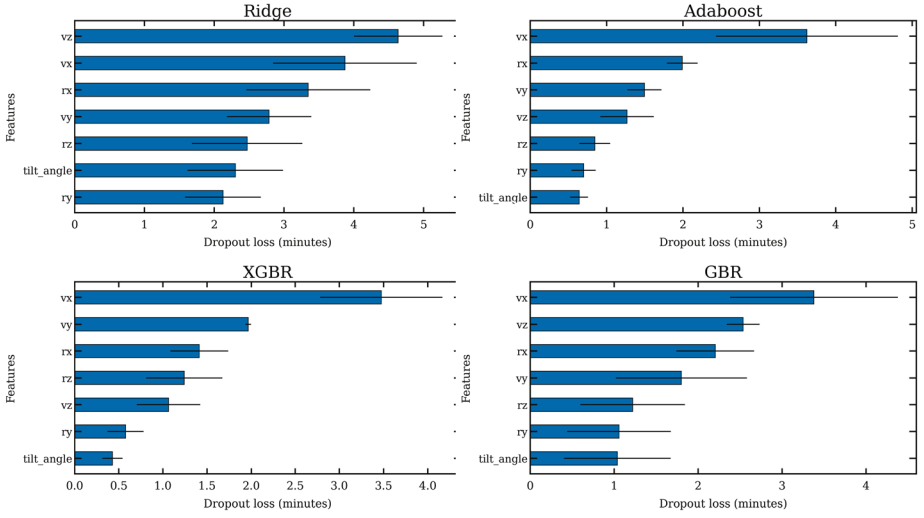


**Figure 7** Feature importance for bin3 where solar-wind speed is greater than 600 km $s^{-1}$. The error bar represents the standard deviation of the propagation-delay error obtained across 2-fold cross-validation for the bin3 dataset.

to be more important than the position of ACE in estimating the propagation time. Also, Earth's dipole tilt feature seems to be least significant in boosting models for high and very high-speed solar wind.

## 4. Conclusion

Eight different ML models were analyzed for their efficacy in determining the propagation delay of solar-wind disturbances upon their observation at Lagrangian point L1. The flat-plane propagation method is operational in real time but exhibits a greater uncertainty. The simplistic assumption of propagation delay is complicated by numerous factors such as the varying position of ACE at L1 and the tilted orientation of the IMF's plane of propagation. Despite numerous empirical methods and MHD models accounting for the tilt of the IMF plane along the parker spiral to minimize the uncertainty, real-time implementation of these model presents challenges. This work attempts to estimate the propagation delay using ML models that include the position of ACE in all three coordinates ($r_x$, $r_y$, $r_z$) and the three vector components of the solar wind ($V_x$, $V_y$, $V_z$). In addition to those input parameters, this work attempts to study the contributions of the Earth's magnetic-axis tilt, which influences the way the solar disturbances interact with the Earth's magnetosphere. Further, we compare each model's accuracy in the estimation of the propagation delay for different ranges of solar-wind speed. Overall, we observe that for slow solar-wind speed, Ridge regression performs well among ML models with an error of 5 min, whereas vector-based delay gave around 4.5 min. For high-speed solar wind, boosting models adaboost, XGBR, and GBR perform better than the vector-based propagation-delay method. We observe that the uncertainty in the propagation delay is greater in slow solar-wind speeds, as it is vulnerable to compression by high-speed solar disturbances along with other factors. Thus, the position of ACE and the plane-orientation angle are important for estimating the propagation delay, whereas the uncertainty is less in high-speed solar disturbances. We assess the significance of input parameters of the model's estimation using an interpretable ML method known as a variable-importance measure. We observe that $V_x$ is the most relevant parameter to the prediction as it aligns with the existing understanding of solar-wind propagation by flat-plane propagation. We also observed that the position of ACE was next relevant in the slow solar-wind speed for the ridge regression model. This relevance may be attributed to the inclination of the propagation front within the ($X$, $Y$)-plane, introducing a dependency on propagation time related to the spacecraft's position (Ridley et al. 1998). In the case of high-speed solar wind, components of the solar-wind vector have more contribution to the estimation of the propagation delay than the position of ACE (Weimer et al. 2002). We observe that the impact of the dipole tilt can be seen only in the slow-speed solar wind, whereas the features are not observed as significant in the high-speed solar wind. Once trained, these models can be implemented in real-time forecasting and can be trained easily with past data. Additional investigation of solar-wind parameters, including solar-wind pressure and magnetic-field components at L1 can be done. Continuous time-series data of these additional solar-wind variables for a given shock event at L1 can be considered as input, which can provide additional information on the nature of the disturbance.

**Acknowledgements** We thank the anonymous reviewer for the constructive comments and suggestions, which greatly improved the readability and quality of the paper. We express gratitude for the availability of ACE data, generously provided by the ACE Science Center at Caltech (www.srl.caltech.edu/ACE). This paper uses data from the Heliospheric Shock Database, generated and maintained at the University of Helsinki to determine the shock list.

**Author contributions** Saurabh Das and Hemapriya Raju conceptualized the idea. Hemapriya Raju contributed to the model development, analysis of the results, it's interpretation, and the initial drafting of the manuscript. Saurabh Das provided supervision throughout the manuscript. Both the authors contributed to refining the final version of the manuscript and reviewed it.

**Data Availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing interests** The authors declare no competing interests.

## References


Awad, M., Khanna, R.: 2015, *Support Vector Regression* **67**, Apress, Berkeley, 978. DOI.

Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: 2021, Dalex: responsible machine learning with interactive explainability and fairness in python. *J. Mach. Learn. Res.* **22**, 1.

Baumann, C., McCloskey, A.E.: 2021, Timing of the solar wind propagation delay between L1 and Earth based on machine learning. *J. Space Weather Space Clim.* **11**, 41. DOI. ADS.

Blanchard, G.T., Bankston, D.: 2002, Improved interplanetary magnetic field propagation timing by correction of the phase front orientation using two spacecraft. *J. Geophys. Res. Space Phys.* **107**, SSH 6. DOI.

Burlaga, L.F., Ogilvie, K.W.: 1969, Causes of sudden commencements and sudden impulses. *J. Geophys. Res.* **74**, 2815. DOI. ADS.

Cameron, T., Jackel, B.: 2016, Quantitative evaluation of solar wind time-shifting methods. *Space Weather* **14**, 973. DOI.

Cameron, T.G., Jackel, B.: 2019, Using a numerical MHD model to improve solar wind time shifting. *Space Weather* **17**, 662. DOI.

Cash, M.D., Witters Hicks, S., Biesecker, D.A., Reinard, A.A., de Koning, C.A., Weimer, D.R.: 2016, Validation of an operational product to determine L1 to Earth propagation time delays. *Space Weather* **14**, 93. DOI.

Collier, M.R., Slavin, J.A., Lepping, R.P., Szabo, A., Ogilvie, K.: 1998, Timing accuracy for the simple planar propagation of magnetic field structures in the solar wind. *Geophys. Res. Lett.* **25**, 2509. DOI.

Crooker, N.U., Siscoe, G.L., Russell, C.T., Smith, E.J.: 1982, Factors controlling degree of correlation between ISEE 1 and ISEE 3 interplanetary magnetic field measurements. *J. Geophys. Res. Space Phys.* **87**, 2224. DOI.

Eggington, J.W.B., Eastwood, J.P., Mejnertsen, L., Desai, R.T., Chittenden, J.P.: 2020, Dipole tilt effect on magnetopause reconnection and the steady-state magnetosphere-ionosphere system: global MHD simulations. *J. Geophys. Res. Space Phys.* **125**, e2019JA027510. DOI.

Friedman, J.H.: 2002, Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367. DOI. Nonlinear Methods and Data Mining.

Fujita, S.: 2019, Response of the magnetosphere–ionosphere system to sudden changes in solar wind dynamic pressure. *Rev. Mod. Plasma Phys.* **3**, 2. DOI.

Gosling, J.T., Asbridge, J.R., Bame, S.J., Hundhausen, A.J., Strong, I.B.: 1967, Discontinuities in the solar wind associated with sudden geomagnetic impulses and storm commencements. *J. Geophys. Res.* **72**, 3357. DOI.

Ho, T.K.: 1995, Random decision forests. In: *Proc. 3rd Int. Conf. Doc. Anal. Recogn.* **1**, 278. DOI.

Horbury, T.S., Burgess, D., Fränz, M., Owen, C.J.: 2001, Prediction of Earth arrival times of interplanetary southward magnetic field turnings. *J. Geophys. Res. Space Phys.* **106**, 30001. DOI.

Joselyn, J.A., Tsurutani, B.T.: 1990, Geomagnetic sudden impulses and storm sudden commencements: a note on terminology. *Eos Trans. AGU* **71**, 1808. DOI.

Karl Laundal, K., Reistad, J., Smith, A., Hovland, A.: 2020. Dipole.

Kelly, T.J., Crooker, N.U., Siscoe, G.L., Russell, C.T., Smith, E.J.: 1986, On the use of a sunward libration-point-orbiting spacecraft as an interplanetary magnetic field monitor for magnetospheric studies. *J. Geophys. Res. Space Phys.* **91**, 5629. DOI.

Lavraud, B., Thomsen, M.F., Borovsky, J.E., Denton, M.H., Pulkkinen, T.I.: 2006, Magnetosphere preconditioning under northward IMF: Evidence from the study of coronal mass ejection and corotating interaction region geoeffectiveness. *J. Geophys. Res. Space Phys.* 111. DOI.

Lepping, R.P., Behannon, K.W.: 1980, Magnetic field directional discontinuities: 1. Minimum variance errors. *J. Geophys. Res. Space Phys.* **85**, 4695. DOI.

Liu, Z.-Q., Lu, J.Y., Kabin, K., Yang, Y.F., Zhao, M.X., Cao, X.: 2012, Dipole tilt control of the magnetopause for southward IMF from global magnetohydrodynamic simulations. *J. Geophys. Res. Space Phys.* 117. DOI.

Lu, J.Y., Liu, Z.-Q., Kabin, K., Jing, H., Zhao, M.X., Wang, Y.: 2013, The IMF dependence of the magnetopause from global MHD simulations. *J. Geophys. Res. Space Phys.* **118**, 3113. DOI.

Mailyan, B., Munteanu, C., Haaland, S.: 2008, What is the best method to calculate the solar wind propagation delay? *Ann. Geophys.* 26. DOI.

Maynard, N.C., Burke, W.J., Sandholt, P.E., Moen, J., Ober, D.M., Lester, M., Weimer, D.R., Egeland, A.: 2001, Observations of simultaneous effects of merging in both hemispheres. *J. Geophys. Res.* **106**, 24551. DOI. ADS.

McComas, D.J., Bame, S.J., Barker, P., Feldman, W.C., Phillips, J.L., Riley, P., Griffee, J.W.: 1998, Solar Wind Electron Proton Alpha Monitor (SWEPAM) for the advanced composition explorer. *Space Sci. Rev.* **86**, 563. DOI.

McDonald, G.C.: 2009, Ridge regression. *WIREs Comput. Stat.* **1**, 93. DOI.

Milan, S.E., Carter, J.A., Bower, G.E., Fleetham, A.L., Anderson, B.J.: 2022, Influence of off-Sun-Earth line distance on the accuracy of L1 solar wind monitoring. *J. Geophys. Res. Space Phys.* **127**, e2021JA030212. DOI.

Munteanu, C., Haaland, S., Mailyan, B., Echim, M., Mursula, K.: 2013, Propagation delay of solar wind discontinuities: comparing different methods and evaluating the effect of wavelet denoising. *J. Geophys. Res. Space Phys.* **118**, 3985. DOI.

Nishida, A., Cahill, J., Laurence, J.: 1964, Sudden impulses in the magnetosphere observed by explorer 12. *J. Geophys. Res.* **69**, 2243. DOI. ADS.

O'Brien, C., Walsh, B.M., Zou, Y., Tasnim, S., Zhang, H., Sibeck, D.G.: 2023, PRIME: a probabilistic neural network approach to solar wind propagation from L1. *Front. Astron. Space Sci.* 10. DOI.

Popescu, M.-C., Balas, V.E., Perescu-Popescu, L., Mastorakis, N.: 2009, Multilayer perceptron and neural networks. *WSEAS Trans. Circuits Syst.* **8**, 579.

Pulkkinen, A., Rastätter, L.: 2009, Minimum variance analysis-based propagation of the solar wind observations: Application to real-time global magnetohydrodynamic simulations. *Space Weather* 7. DOI.

Ridley, A.J.: 2000, Estimations of the uncertainty in timing the relationship between magnetospheric and solar wind processes. *J. Atmos. Solar-Terr. Phys.* **62**, 757. DOI.

Ridley, A.J., Lu, G., Clauer, C.R., Papitashvili, V.O.: 1998, A statistical study of the ionospheric convection response to changing interplanetary magnetic field conditions using the assimilative mapping of ionospheric electrodynamics technique. *J. Geophys. Res. Space Phys.* **103**, 4023. DOI.

Russell, C.T., Siscoe, G.L., Smith, E.J.: 1980, Comparison of ISEE-1 and −3 interplanetary magnetic field observations. *Geophys. Res. Lett.* **7**, 381. DOI.

Schwartz, S.J.: 1998, Analysis methods for multi-spacecraft data. *ISSI Sci. Rep. Ser.* **1**, 249.

Sonnerup, B.U.Ö., Scheible, M.: 1998, Minimum and maximum variance analysis. *ISSI Sci. Rep. Ser.* **1**, 185. ADS.

Tsurutani, B., Baker, D.: 1979, Substorm warnings: an ISEE-3 real time data system. *Eos Trans. AGU* **60**, 701. DOI.

Weimer, D.R., King, J.H.: 2008, Improved calculations of interplanetary magnetic field phase front angles and propagation time delays. *J. Geophys. Res. Space Phys.* 113. DOI.

Weimer, D.R., Ober, D.M., Maynard, N.C., Burke, W.J., Collier, M.R., McComas, D.J., Ness, N.F., Smith, C.W.: 2002, Variable time delays in the propagation of the interplanetary magnetic field. *J. Geophys. Res. Space Phys.* **107**, 1210. DOI. ADS.

Weimer, D.R., Ober, D.M., Maynard, N.C., Collier, M.R., McComas, D.J., Ness, N.F., Smith, C.W., Watermann, J.: 2003, Predicting interplanetary magnetic field (IMF) propagation delay times using the minimum variance technique. *J. Geophys. Res. Space Phys.* 108. DOI.

Zhu, C.B., Zhang, H., Ge, Y.S., Pu, Z.Y., Liu, W.L., Wan, W.X., Liu, L.B., Chen, Y.D., Le, H.J., Wang, Y.F.: 2015, Dipole tilt angle effect on magnetic reconnection locations on the magnetopause. *J. Geophys. Res. Space Phys.* **120**, 5344. DOI.