JSWSC

Available online at:
www.swsc-journal.org

RESEARCH ARTICLE   OPEN ACCESS

A catalogue of observed geo-effective CME/ICME characteristics

Ronish Mugatwala[1,2,*], Simone Chierichini[1,2], Gregoire Francisco[1,3], Gianluca Napoletano[1,4],
Raffaello Foldes[4,5], Luca Giovannelli[1], Giancarlo De Gasperis[6], Enrico Camporeale[7,8],
Robertus Erdélyi[2,9,10], and Dario Del Moro[1]

[1] Department of Physics, University of Rome "Tor Vergata", Via della Ricerca Scientifica 1, I-00133 Rome, Italy
[2] SP2RC, School of Mathematics and Statistics, University of Sheffield, Hicks Building, 10 Broomhall, Sheffield S3 7RH, England
[3] Institute of Astrophysics and Space Science, University of Coimbra, 3004-531 Coimbra, Portugal
[4] Dipartimento di Scienze Fisiche e Chimiche, Università dell'Aquila, Via Vetoio, 67100 L'Aquila, Italy
[5] Ecole Centrale de Lyon, CNRS, Universite Claude Bernard Lyon 1, INSA Lyon, LMFA, UMR5509, 69130 Ecully, France
[6] Dipartimento di Fisica, Sapienza Università di Roma, P. le A. Moro 2, Roma, Italy
[7] Space Weather Research Technology Education Center (SWx-TREC), University of Colorado, Boulder, CO 80309, USA
[8] Queen Mary University of London, London E1 4NS, UK
[9] Department of Astronomy, Eötvös Loránd University, Pázmány Péter sétány 1/A, H-1112 Budapest, Hungary, Budapest, Hungary
[10] Gyula Bay Zoltán Solar Observatory (GSO), Hungarian Solar Physics Foundation, Petőfi tér 3, 24, H-5700 Gyula, Hungary

Abstract – One of the goals of Space Weather studies is to achieve a better understanding of impulsive phenomena, such as Coronal Mass Ejections (CMEs), to improve our ability to forecast their propagation characteristics and mitigate the risks to our technologically driven society. The essential part of achieving this goal is to assess the performance of forecasting models. To this end, the quality and availability of suitable data are of paramount importance. In this work, we merged publicly available data of CMEs from both in-situ and remote observations in order to build a dataset of CME properties. To evaluate the accuracy of the dataset and confirm the relationship between in-situ and remote observations, we have employed the Drag-Based Model (DBM) due to its simplicity and modest consumption of computational resources. In this study, we have also explored the parameter space for the drag parameter and solar wind speed using a Monte Carlo approach to evaluate how efficiently the DBM determines the propagation of CMEs for the events in the dataset. The geoeffective CMEs selected as a result of this work are compliant with the hypothesis of DBM (isolated CME, constant solar wind speed beyond 20 $R_\odot$) and also yield further insight into CME features such as arrival time and arrival speed at L1 point, lift-off time, speed at 20 $R_\odot$ and other similar quantities. Our analysis based on the acceptance rate in the DBM inversion procedure shows that almost 50% of the CME events in the dataset are well described by DBM as they propagate in the heliosphere. The dataset includes statistical metrics for the DBM model parameters. The probability distribution functions (PDFs) for the free parameters of DBM have been derived through a Monte Carlo-like inversion procedure. Probability distribution functions obtained from this work are comparable to PDFs employed in previous works. The analysis showed that there exist two different most probable values (median values) of solar wind speed for DBM input based on slow ($w_{slow} \approx 386$ km/s) and fast ($w_{fast} \approx 547$ km/s) solar wind type. The most probable value for the drag parameter ($\gamma \approx 0.687 \times 10^{-7}$ km^{-1}) in our study is somewhat higher than the values reported in previous studies. Using a data-driven approach, this procedure allows us to present a homogeneous, reliable, and robust dataset for the investigation of CME propagation. Additionally, possible CME events are identified where the DBM prediction is not valid due to model limitations and higher uncertainties in the input parameters. These events require further thorough investigation in the future.

Keywords: Coronal Mass Ejection / Heliosphere / Magnetohydrodynamic Drag / Space Weather / Forecast tools

*Corresponding author: ronish@roma2.infn.it

1 Introduction

ICMEs (Interplanetary Coronal Mass Ejections) are eruptions of plasma and magnetic fields from the Sun's corona that propagate in the Heliosphere (Webb & Howard, 2012). These plasma and magnetic field structures are ejected from the Sun, travel through the interplanetary space environment, and reach the 1 AU range within 1–5 days (Chen, 2011). In in-situ data, ICMEs can be discerned from the average solar wind by their distinct signatures, such as the enhanced magnetic field, the higher particle speed, and the variations in plasma density (Liu et al., 2010; Papaioannou et al., 2016). They can also be observed remotely by using instruments such as coronagraphs (particularly SOHO/LASCO with coronagraphs C1/C2/C3 Brueckner et al., 1995; Domingo et al., 1995, STEREO/SECCHI with COR1/COR2 Kaiser et al., 2008; Howard et al., 2008), and Heliographic imagers (HI1/HI2) (Eyles et al., 2009).

ICMEs are among the main drivers of Space Weather, impacting the space environment and human technologies (Tsurutani et al., 1988; Gosling et al., 1991; Schwenn, 2006; Pulkkinen, 2007; Temmer, 2021). The plasma and magnetic fields ejected from the Sun can interact with Earth's magnetic field, leading to geospace disturbances (Koskinen & Huttunen, 2007), which affect a wide range of technological systems in space, such as satellites, telecommunications, and the GNSS systems (Shea & Smart, 1998; Schrijver & Siscoe, 2010; Aquino & Sreeja, 2013; Piersanti et al., 2017). The present strategies to mitigate the effects of ICMEs on space-based technologies and infrastructures require the knowledge of the ICME arrival time with low uncertainty to allow operators to take action to protect their equipment, by shutting them down or putting them in a safe mode (Barbieri & Mahmot, 2004; Sreeja, 2016; Veettil et al., 2019).

In the last decades, space agencies have designed and launched missions to observe the Sun and monitor the solar wind characteristics, and track CMEs and ICMEs as they travel through space, with the aim to study their interactions with the interplanetary environment. Despite these advancements in space weather forecasting, accurately predicting the characteristics of ICMEs such as their Time-of-Arrival (ToA) and Speed-at-Arrival (SaA) at Earth, as well as the magnitude and direction of the southward component of their magnetic field (which is crucial for determining the intensity of geomagnetic storms Koskinen & Huttunen, 2006), remains a challenging task for the scientific community (Manchester et al., 2017; Riley et al., 2018; Vourlidas et al., 2019).

Following the evolution of numerical methods and the increase of available computational power, a number of empirical methods, physics-based analytical models, and MHD numerical simulations for the ICME kinematics have been developed. In the MHD approximation, the boundary conditions are derived from observed magnetograms and coronographic images and model the propagation of the ejecta by numerically solving the magneto- hydrodynamic equations (ENLIL, HAFv.2 (Hakamada-Akasofu-Fry version 2)+3DMHD, EUHFORIA (EUropean Heliospheric FORecasting Information Asset) Odstrcil et al., 2003; Wu et al., 2007; Pomoell & Poedts, 2018). These simulations allow for the inclusion and consideration of the physical processes being modelled. However, their use requires substantial computing resources due to their computationally intensive nature, making them expensive to run. The complete understanding of the physical processes involved in the Sun-Earth relation relies heavily on numerical modelling techniques. However, with present observation capabilities, the forecasting performance of empirical and analytical methods are comparable to, or in some cases slightly better than, those achieved with numerical methods due to uncertainties in the input parameters (Manchester et al., 2017). This implies that the existing empirical and analytical approaches are still effective and competitive in terms of their predictive capabilities and that the near future of space weather forecasting lies with the use of these computationally light approaches and Machine Learning (ML).

In general, analytical methods are computationally lighter and their parameters can be easily updated with new incoming data. Also, physics-based analytical models (e.g. Vršnak et al., 2013; Rollett et al., 2016; Paouris & Mavromichalaki, 2017; Napoletano et al., 2018) can shed light on the ICME dynamics, and this knowledge would possibly help us in refining also numerical methods. On the other hand, the relationships between ToA and SaA and various CME parameters measured at (or close to) their launch, have been used in empirical prediction methods (e.g. Manoharan, 2006; Gopalswamy, 2009), and most recently in a plethora of ML approaches. ML techniques have become more and more used in space weather, as recently reviewed in Camporeale (2019). In the last years, there have been many attempts to leverage ML algorithms to obtain the characteristics of an ICME at L1 from the associated CME observables (Bobra & Ilonidis, 2016; Liu et al., 2018; Wang et al., 2019, just to list a few). These ML algorithms use catalogues of CME/ICME characteristics for the training, in order to set their parameters, validate their results and check their performances. Consequently, it becomes more and more important to build CME/ICME datasets with a large number of events and small uncertainties (ML methods typically need numerous, relevant and reliable examples in the datasets in order to give accurate results VanderPlas et al., 2012; Ivezić et al., 2014).

In this paper, we present a method to update the catalogue of CME–ICME pairs published in Napoletano et al. (2022), by using a constrained Monte-Carlo strategy to validate its entries. The constrained Monte Carlo strategy allowed us to explore the parameter space in a more effective way. We then make use of this updated catalogue to revisit the Probability Distribution Functions (PDFs) to use for the P-DBM method (Napoletano et al., 2018; Del Moro et al., 2019). Finally, we present a comparison of these PDFs for different solar wind conditions and against previous literature.

The paper is organized as follows. Section 2 describes the DBM model and the mathematical methodologies to retrieve PDFs from the catalogue. In Section 3, we analyze the results of the inversion procedure and use them to relabel the CME/ICME catalogue entries and to obtain PDFs for different ICME types. Section 4 is dedicated to conclusions and discussions. The CME–ICME dataset compiled and used in this work can be found at https://zenodo.org/record/8063404 and a description of the different column headers is provided in Appendix A.

2 Methods

2.1 Drag-Based Model (DBM)

The Drag-Based Model is one of the simplest models that describes CME propagation through the heliosphere. Due to

its simplicity and calculation speed, it is one of the most popular models used in CME forecast tools. In recent years, DBM has been used in many studies to describe CME propagation which is summarised in Dumbović et al. (2021). DBM is based on the assumption that the Lorentz force responsible for CME launch is negligible in the upper part of the solar corona, specifically beyond a heliocentric distance of 20 $R_\odot$. However, this assumption is not always valid; for many CME events the Lorentz force is still comparable with the drag force and the exact distance can vary from event to even Vršnak (2001), Vršnak et al. (2004), Sachdeva et al. (2015, 2017). Beyond this heliocentric distance, the dynamics of the ICME are dominantly governed by its interaction with the ambient solar wind via MHD drag (Cargill, 2004; Vršnak et al., 2013). Due to MHD drag force ICMEs that are faster (slower) than solar wind tend to decelerate (accelerate) during propagation, which was also supported by observations (Gopalswamy et al., 2000). CME radial acceleration according to the DBM approach is given as:

$$a(r) = -\gamma(v(r)_{CME} - w)|v(r)_{CME} - w|, \quad (1)$$

where $a(r)$ and $v(r)_{CME}$ are the instantaneous acceleration and speed of ICME, respectively, w is the instantaneous ambient solar wind speed, and γ is the drag parameter that is also called drag efficiency. It is important to note that all the quantities in equation (1) are space and time-dependent. Also, beyond 20 $R_\odot$, γ and w may be approximated to be constant throughout the heliosphere (Cargill, 2004; Vršnak et al., 2013). Under such approximation, equation (1) can be solved analytically to obtain heliospheric distance and speed of ICME as a function of time (Vršnak et al., 2013):

$$v(t) = \frac{v_0 - w}{1 \pm \gamma(v_0 - w)t} + w, \quad (2)$$

$$r(t) = \pm\frac{1}{\gamma}\ln(1 \pm \gamma(v_0 - w)t) + wt + r_0, \quad (3)$$

where $\pm$ sign accounts for accelerated/decelerated CMEs i.e., plus for $v_0 > w$ and minus for $v_0 < w$. Equations (2) and (3) give us the speed and distance as a function of CME propagation time from an initial distance (at $t = 0$) r_0 and take-off speed v_0. From those, one can determine the transit time t_{1AU} and impact speed v_{1AU} at 1AU.

2.2 DBM inversion procedure

DBM solution, as given in Vršnak et al. (2013), can be used to obtain the analytical values of free DBM parameters. If the ICME follows the DBM model, and if its boundary conditions, i.e, initial position r_0, initial speed v_0, ToA t_{1AU} and impact speed v_{1AU} are known, then the free parameters of the model, namely drag parameter γ and solar wind speed w, can be obtained via a mathematical inversion of the set of presented equations (2) and (3).

$$\frac{(v_0 - w)(v_{1AU} - w)t_{1AU}}{(v_0 - v_{1AU})}\ln\left[\frac{(v_0 - v_{1AU})}{(v_{1AU} - w)} + 1\right]$$
$$+ wt_{1AU} + r_0 - r_{1AU} = 0. \quad (4)$$

Equation (4) is solved numerically to obtain w, then equation (5) is used to directly compute γ:

$$\gamma = \frac{(v_0 - v_{1AU})}{(v_0 - w)(v_{1AU} - w)t_{1AU}}. \quad (5)$$

2.3 Mathematical framework

In search for the unique distribution for the free DBM parameters, we applied the DBM inversion procedure to the existing dataset, published in the previous works of Napoletano et al. (2018, 2022). A comprehensive description of the dataset is provided in Appendix A, while the summary of a few particular quantities used in this study and associated results are tabulated in Table A.1. In the process of DBM inversion, we discovered that the majority of the CME events in the dataset lack analytical solutions for equations (4) and (5). This collided with our null hypothesis for the dataset: that the DBM assumptions (Vršnak et al., 2013) were generally valid for the propagation of most of the CME events in the list and that the large experimental uncertainties associated with the ICME quantities allowed for a parameter space large enough to find at least one γ-w couple to solve the equation system. The reason behind this discrepancy is that errors associated with the initial position (r_0), target position (r_1), transit time (t_{1AU}), impact speed (v_1) and initial speed (v_0) were omitted in the inversion procedure. An alternative explanation could be that DBM is not properly describing the CME motion (e.g. w = constant is not a realistic approximation; CME–CME interaction is also possible).

However, for this study, we adhere to our null hypothesis and consider the possibility of including uncertainties for the measured CME properties. To incorporate the errors associated with those quantities, we adapted a pairwise selection approach. It is important to highlight that Napoletano et al. (2022) also employed a probabilistic approach in the inversion procedure to obtain w and γ. In order to do that, they assumed that $[r_0, r_1, v_0, v_1, t_{1AU}]$ follows a normal distribution and draws random samples, where the majority of samples are concentrated around the peak of the Gaussian curve. However, our pairwise approach allowed us to explore other parts of parameter space where less probable values exist. We have assumed that two parameters, r_0 and r_1, do not suffer any errors because their values are fixed. We took $r_0 = 20$ $R_\odot$ and, for r_1 we have used the actual Sun-Earth Distance at a time when CME is at r_0. The arrival speed of CME in the dataset is calculated as a mean of solar wind speed during a disturbance in plasma and therefore it has an associated intrinsic error. The error associated with arrival speed is relatively small compared to the initial speed and arrival time, therefore it is neglected in the study. The two remaining quantities with large errors are thus v_0 and t_{1AU}. Next, we made a pairwise selection of (t_{1AU}, v_0) for each DBM inversion iteration from the normal distribution followed by both quantities where μ is the observed value ("Transit_Time" and "v_r" is taken as μ) and σ is an error associated with the observed quantity ("Transit_time_err" and "v_r_err" taken as σ). Here "v_r" is the deprojected speed of CME at 20 $R_\odot$ taken from the Napoletano et al. (2022) and "v_r_err" is an error associated with deprojected speed. It is important to keep in mind that the tails of the normal distribution function are 3σ width. For a pairwise selection, we draw 200 samples for t_{1AU} and the same for v_0. So in the end, we have a total of 40,000 possible pairs. After this pair selection, we performed the DBM inversion to obtain values of w and γ, respectively.
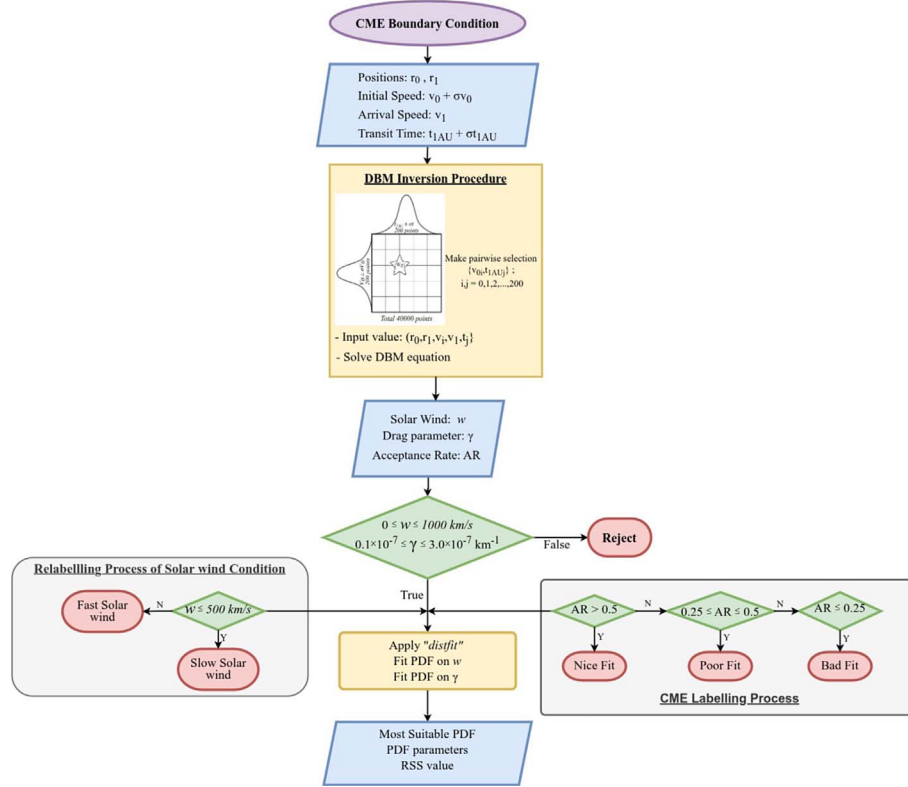
Figure 1. Schematic of DBM inversion procedure. The values of boundary conditions are fed into the equations (4) and (5) using a pairwise approach to obtain w and γ. The obtained values are checked for selection criteria. The accepted values are used to determine the solar wind condition, the most suitable PDF of model parameters and the CME labelling scheme.

The DBM inversion procedure is a Monte Carlo process and after the inversion procedure, we have 40,000 possible solutions for w and γ. Many of these values can not be physically feasible, for example, negative values of w. Vršnak et al. (2013) provided a brief description of γ in their work, and from there we deduced that the drag parameter γ has a relation with the mass and cross-sectional area of CME along with ambient solar wind density. In our primary analysis, we found that the inversion procedure also provides very high values for γ which can not be explained by the typical value range of CME features like mass, cross-sectional area and background solar wind density. Therefore, it is necessary to employ constraints on the values obtained through the inversion procedure. The constraints that we imposed on inversion values are given below.

1. $0 \leq w \leq 1000$ km/s
   Solar wind speed cannot be negative and the typical speed for fast solar wind in literature is 800 km/s. It is worth noting that the condition of realistic solar wind speed in Paouris et al. (2021) is 300–600 km/s which is very narrow compared to us.
2. $0.1 \times 10^{-7} \leq \gamma \leq 3.0 \times 10^{-7}$ km^{-1}
   It is important to note that the typical range for the γ parameter, in Vršnak et al. (2013), is $0.2 - 2.0 \times 10^{-7}$, but we widen this range to accept a few more extreme solutions. Similarly, Paouris et al. (2021) has a range of $0.01 - 0.59 \times 10^{-7}$ for realistic drag parameter but their obtained values are in the range of $0.21 - 0.42 \times 10^{-7}$

(see Table 4 of Paouris et al., 2021) which is comparable to our range.

After this, we derive the four main quantities namely W_{mean}, γ_{mean}, W_{opt} and γ_{opt} from the accepted values of w and γ; the opt values correspond to the DBM input that produced the minimum deviation from the observed transit time. In order to evaluate the "goodness" of the inversion procedure, we define the "Acceptance Rate" as the ratio between the number of meaningful solutions to the total number of possible solutions, represented by the number of samples.

$$\text{Acceptance Rate (AR)} = \frac{\text{no. of physically feasible solutions}}{\text{Total no. of solutions}}$$

$$= \frac{m}{n \times n}. \tag{6}$$

Here, m is the no. of solutions accepted after applying constrain and n is no. the of samples drawn from the $t_{1\text{AU}}$ and v_0 distributions each. Figure 1 shows the flow diagram for the DBM inversion procedure that we implemented on the CME dataset and how the results of the inversion procedure are analysed.

3 Results

3.1 Inversion procedure results

The inversion procedure was performed on the entire CME–ICME pair catalogue and it turned out to be successful for 204
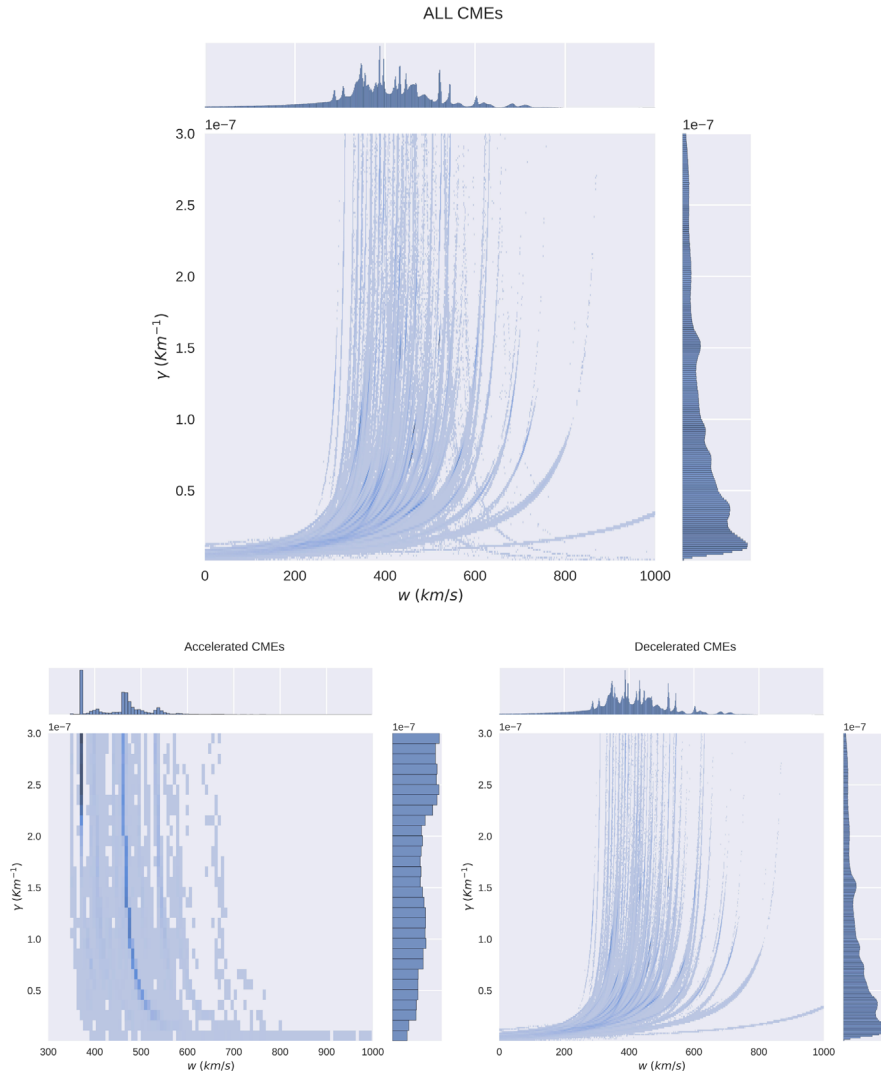
Figure 2. Joint distribution of (γ,w) from the inversion procedure. Top Panel: (γ,w) Phase space for the whole dataset (3,644,748 values). Bottom left Panel: (γ,w) phase space for the dataset of accelerated CMEs (25,428 values). Bottom right Panel: (γ,w) phase space for the dataset of decelerated CMEs (3,619,320 values).

out of 213 events. At the end of the inversion procedure, we obtained 3,664,748 possible values of w and γ that enable us to provide a statistical distribution for them. Figure 2 illustrates, the (γ,w) phase space for the entire ICME dataset and there we can identify the predictive line for a few individual CMEs. From the DBM equation (1), one can easily notice that CMEs either accelerate or decelerate during their propagation. Based on these propagation conditions, we derived two different distributions divided into accelerated and decelerated CMEs. Furthermore, a free DBM parameter w can also be divided into two groups called the slow and fast solar wind, and therefore we can draw two more joint distributions based on solar wind speed conditions.

3.2 Determining the quality of inversion process for each CME event

It is important to note that, we claimed that the DBM inversion was successful for 204 events and therefore there should be 8,160,000 possible values of γ and w, which are more than double the numbers we have obtained from the DBM inversion procedure. This discrepancy results from either the DBM inversion procedure failing or the obtained values of (γ,w) being discarded as they did not fulfil the constraints. This can also be observed in the (γ,w) phase space of different CME events. Based on the density in the (γ,w) phase space, we label the event as "Optimal Fit", "Suboptimal Fit" and "Inadequate Fit". This labelling helps us to determine if the propagation of CME events in the dataset can be described by DBM. To stay consistent in the labelling procedure, we used the Acceptance Rate (AR) defined by equation (6). The description for the labels is as follows.

1. Optimal (Nice) Fit: AR > 0.5; the DBM approximation is demonstrably accurate for this kind of CME event as the inversion procedure is successful for more than 50% of the pairs. Therefore, there is a very sharp trendline in (γ,w) phase space.
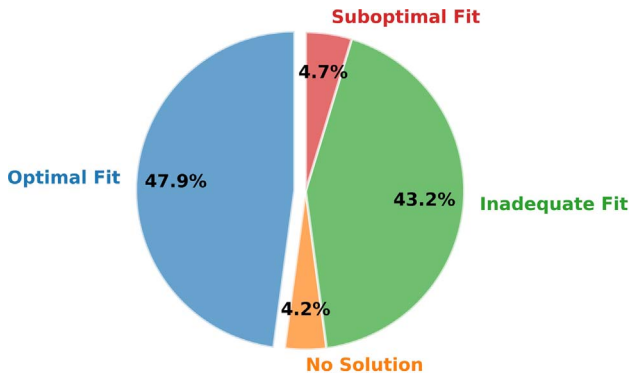
Figure 3. A pie chart showing a percentage of events in the "Optimal Fit", "Suboptimal Fit" and "Inadequate Fit".

2. Suboptimal (Poor) Fit: $0.25 \leq AR \leq 0.5$; the DBM is moderately accurate as one can still see the trendline in (γ, w) phase space.
3. Inadequate (Bad) Fit: $AR < 0.25$; the DBM approximation is less applicable for the events and it is hard to find the trend line in phase space.

Figure 3 shows the percentage of events in each assigned label. We want to stress here the fact that, this labelling scheme is a key point for the dataset that is created as a result of this work. This labelling helps us to determine which CME event in the dataset is described well by the DBM. Events flagged as "Suboptimal Fit" or "Inadequate Fit" require further investigation. Hereafter, we only focused on the *Optimal Fit* events to obtain the PDFs for w and γ as it helps to improve the PDFs. Eventually, these better statistics will lead to better accuracy in CME arrival forecasting.

3.3 Relabelling the solar wind condition

We found that there are only 28 CME events that are accelerating during propagation, these are around 13% of the entire dataset, therefore statistics for accelerating CMEs are not very well resolved. In order to find a distribution for the free DBM parameters we established a group of CME based on solar wind conditions. A dataset that is already obtained as a part of previous work of Napoletano et al. (2022) contains information about the solar wind speed type (see Appendix A Column: SW_type -S/F) based on the presence of Coronal Holes close to the source of a CME. The group of CMEs formed based on coronal hole presence data provides two completely overlapping distributions for fast and slow solar wind speeds. This difference is inconsistent with our expected knowledge of distinct solar and fast solar wind circumstances. The discrepancy arises from incorporating the presence of coronal holes in a calculation of PDF for solar wind speed without explicit consideration of solar wind speed itself. For many events, this connection between solar wind type and a coronal hole is wrong due to various reasons such as CME does not encounter a fast solar wind stream at all during propagation. This leads to the overlap of distribution, as DBM inversion provides a small value of w and the presence of coronal holes identifies that solar wind as fast solar wind. Furthermore, the large standard deviation makes

the model unsuitable for precise and reliable real-time space weather forecasting applications. Therefore we redefine the solar wind type associated with each CME by threshold $W_{sim} \geq 500$ km/s to discriminate the fast solar wind from the slow one. This threshold is similar to one that is used in Napoletano et al. (2018). In Figure 4 (γ, w) phase space is shown for the two "SW_type" and "Wind_type" solar wind labelling. One important point to note here is, that the tail part of any distribution in a negative region is due to the plotting style not due to the presence of any value. Also, from now onward we focus on this new labelling scheme for solar wind speed.

3.4 PDF for solar wind speed

From the joint distribution shown in Figure 2, we can extract a distribution function for the solar wind speed w. Here, we have fitted Gaussian, Student-t and Lognormal functions to the distribution function as these three functions returned a better fit among different PDFs available in the *distfit* package (Taskesen, 2023). In Figure 5, the histogram obtained from the dataset and fitted PDFs are shown. Here we have considered the RSS (Residual Sum of Squares) value to determine which one is the best fit. In most cases, all 3 distribution functions show a similar RSS value which is clear from the figure as well. So, in the end, we decided to select the Gaussian distribution function for the solar wind speed w to be consistent with previous works, e.g., Dumbović et al. (2018, 2021), Napoletano et al. (2018, 2022).

We categorized our dataset into Slow and Fast CMEs based on the ambient solar wind condition experienced by the CME during its propagation (using a threshold of 500 km/s to separate fast and slow solar wind conditions), as described before in this section, and attempted again to fit the same three distribution functions. Unlike the prior attempt, the fitting's RSS value is not the same for slow and fast solar wind conditions. For slow CMEs, the "student-t" distribution describes the best PDF while for fast CMEs the lognormal function is the most suitable PDF. Here, we only emphasize the fact that "student-t" and "lognormal" distributions are the best fits and are strongly biased by the hard thresholding. In Figure 6 PDFs for the slow and fast solar wind conditions are shown. The parameters for the fitted distributions are reported in Table 1.

Paouris et al. (2021) studied the same 16 CME–ICME events from Dumbović et al. (2018) to compare the performance of the Effective Acceleration Model (EAM) with Drag Based Ensemble Model (DBEM). They have also performed the inversion technique to find optimal values of solar wind speed w and drag parameter γ. In Table 2, optimal values of w from different studies have been shown. It is important to note that the sample size employed in Napoletano et al. (2022) and this work is large, which helps to explain the higher value of the standard deviation.

3.5 PDF for drag parameter

For the drag parameters, we employed the same methods and distribution functions that we have used for the solar wind to infer the PDF. The RSS values obtained from the various fits are significantly different. The lognormal distribution consistently emerges as the best fit among various considered distribution functions throughout a wide range of cases. In Figure 7, distributions fitting for the accelerated and decelerated CMEs
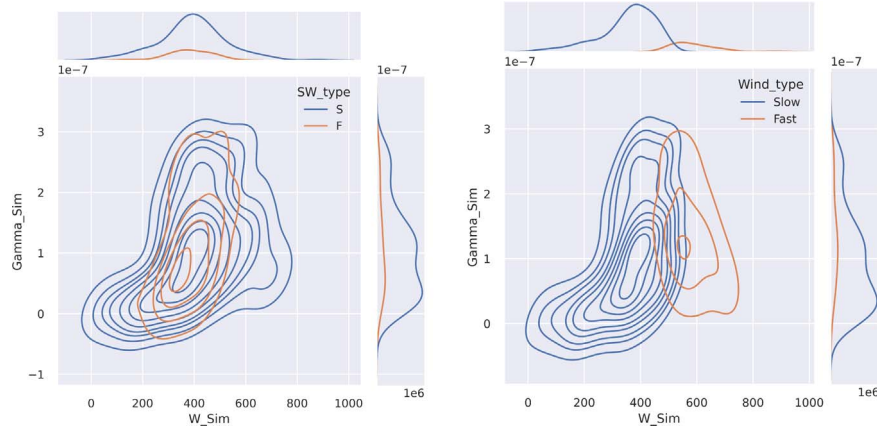
Figure 4. (γ, w) Phase space in different solar wind speed condition labeling scheme. On the x-axis, W_sim shows solar wind speed obtained from a DBM inversion with a unit of km/s while on the y-axis drag (Gamma_sim) value obtained from the inversion procedure is shown on a unit scale of km^{-1}.
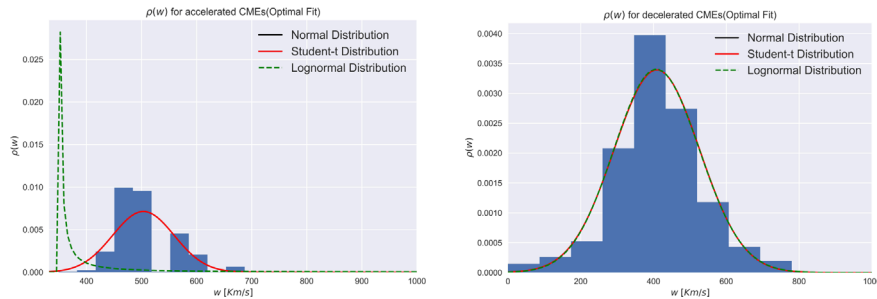
Figure 5. Probability distribution functions for solar wind speed w for accelerated and decelerated CMEs with a kernel density ρ on the y-axis. Left: w PDFs for accelerated CMEs with Optimal Fit label. Note that the normal and student-t distributions overlap with each other. Right: w PDFs for decelerated CMEs with Optimal Fit label. All three distribution functions overlap with each other. The overlapping of functions is evident through RSS values.
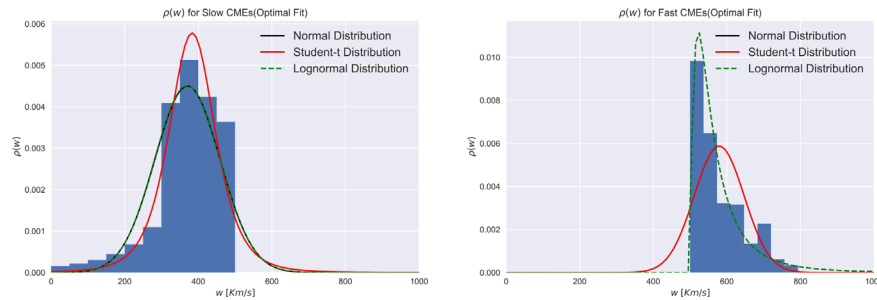
Figure 6. Probability distribution functions for solar wind speed w for slow and fast CMEs. Left: w PDFs for slow CMEs with Optimal Fit label. Right: w PDFs for fast CMEs with Optimal Fit label. (In both cases the Normal and Student-t functions overlapped with each other).

are shown, while in Figure 8 PDFs for slow and fast CMEs are shown. The fitting parameters for the different distribution functions of the drag parameter γ are tabulated in Table 3.

4 Discussion and conclusions

The CME–ICME pair published in Napoletano et al. (2022) is improved by the inclusion of the predicted DBM data, PDF fitting parameters, and various other significant variables e.g., CME arrival time and speed, *Dst* index, source location, B_z component, etc related to each CME–ICME occurrence. By quantifying the success rate of the DBM inversion procedure, we were able to identify a subset of CME–ICME pairs that are well described by the DBM during their heliospheric propagation and added to the dataset. For the space weather community, this kind of categorization can provide significant insight into the conditions that make the DBM forecast fail to predict the correct transit time for a CME event. On the other hand, for those CME events where the DBM forecast is accurate,

Table 1. Parameters for the different functions used to model the solar wind speed distribution. For the Lognormal function, tabulated values can not be used directly as average and standard deviation. The transformation from the fitting parameters to values used in the model can be done by equation (B.4).

CME group	PDF	$\bar{w}$ (km/s)	σ_w (km/s)	Args	RSS
Accelerated	Normal	503.356	55.848	–	0.000538
	Student's-t	503.356	55.848	1.526×10^6	0.000538
	Lognormal	−0.407	3.910	349.009	0.001469
Decelerated	Normal	409.168	117.545	–	0.00046
	Student's-t	409.168	117.543	1.479×10^5	0.00046
	Lognormal	9.524	0.009	$−1.328 \times 10^4$	0.000458
Slow	Normal	370.530	88.585	–	0.000714
	Student's-t	383.169	64.944	4.101	0.000622
	Lognormal	9.784	0.005	$−1.738 \times 10^4$	0.00072
Fast	Normal	579.058	67.871	–	0.002862
	Student's-t	579.058	67.872	1.837×10^6	0.002862
	Lognormal	4.084	0.883	494.597	0.001934

Table 2. Optimal (mean) values for solar wind speed w from different studies.

|	Optimal solar wind speed w (km/s)	Standard deviation σ_w (km/s)
Dumbović et al. (2018)	350	50
Napoletano et al. (2018) (slow)	400	33
Napoletano et al. (2018) (fast)	600	66
Paouris et al. (2021)	431	57
Čalogović et al. (2021)	453	64
Napoletano et al. (2022) (slow)	370	80
Napoletano et al. (2022) (fast)	490	100
This work (slow)	370	88
This work (fast)	579	68
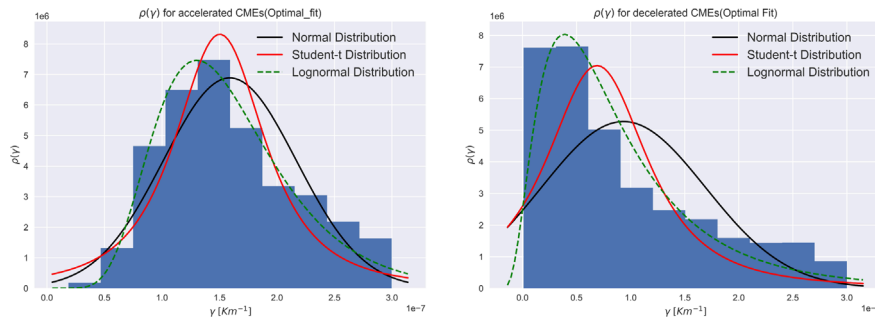
Figure 7. Probability distribution functions for drag parameters γ for accelerated and decelerated CMEs. Left: w PDFs for accelerated CMEs with Optimal Fit label. Right: w PDFs for decelerated CMEs with Optimal Fit label.

it can contribute to providing information about the model parameters w and γ. Thus, all the CME–ICME entries that do not follow the DBM approximation deserve even further investigation, since we cannot tell if a "no solution", a "suboptimal" or an "inadequate" label comes from a possible error in the initial CME–ICME association, a shortage in the geometrical description of the ICME, or something happening during the ICME propagation that cannot be described by the DBM (e.g. a CME–CME interaction). This, however, would require a thorough analysis of every single ICME and is beyond the scope of this work and may be the subject of a different work. The revised CME–ICME collection we are presenting also includes additional details such as the solar wind speed

conditions experienced by propagating CME events, more parameters about the validation of the DBM hypothesis, and information about the acceleration or deceleration mechanisms during their propagation. The list of the improvements over the previous version published by Napoletano et al. (2022) is summarised in Table 5. The revised dataset compiled and used in this work has been published at https://zenodo.org/record/8063404 and a description of its columns is also provided in Appendix A.

Just as mentioned, the subset of events where the DBM approximation holds can be employed to extract the γ and w parameters of the DBM via a Monte Carlo-like inversion procedure. In this statistical study, we consider the uncertainties
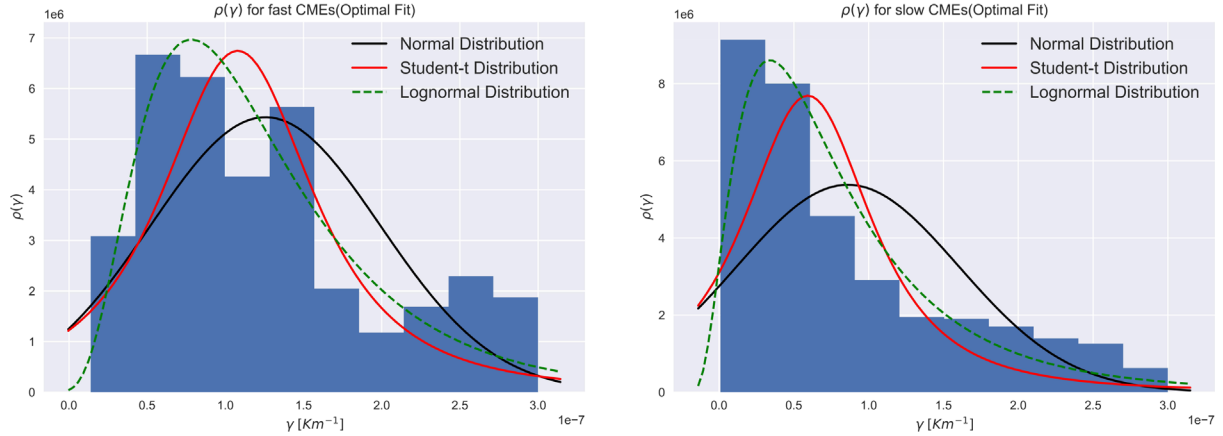
Figure 8. Probability distribution functions for drag parameters γ for slow and fast CMEs. Left: w PDFs for fast CMEs with Optimal Fit label. Right: w PDFs for slow CMEs with Optimal Fit label.

Table 3. Parameters for different PDFs used to model drag parameter distribution. For the Lognormal function, tabulated values can not be used directly as average and standard deviation. The transformation from the fitting parameters to values used in the model can be done by equation (B.4).

CME group	PDF	$\bar{\gamma}$ (km^{-1})	σ_γ (km^{-1})	Args	RSS
Accelerated	Normal	1.590×10^{-7}	5.793×10^{-8}	–	2.528×10^{13}
	Student's-t	1.503×10^{-7}	4.247×10^{-8}	1.988	2.490×10^{13}
	Lognormal	-15.642	0.354	-1.186×10^{-8}	6.385×10^{12}
Decelerated	Normal	9.339×10^{-08}	7.562×10^{-08}	–	1.089×10^{15}
	Student's-t	6.899×10^{-08}	5.016×10^{-08}	1.988	8.029×10^{14}
	Lognormal	-16.178	0.652	0.6518	2.723×10^{14}
Slow	Normal	8.609×10^{-08}	7.419×10^{-08}	–	1.519×10^{15}
	Student's-t	5.936×10^{-08}	4.595×10^{-08}	1.988	1.010×10^{15}
	Lognormal	-16.252	0.658	-2.276×10^{-08}	4.034×10^{14}
Fast	Normal	1.256×10^{-07}	7.342×10^{-08}	–	5.319×10^{14}
	Student's-t	1.079×10^{-07}	5.238×10^{-08}	1.988	4.749×10^{14}
	Lognormal	-15.884	0.518	-1.838×10^{-08}	2.575×10^{14}

associated with the measure and the observation and incorporate them as input for the model and we only consider those CME events with more than 50% acceptance rate in the inversion procedure. The reason behind this criterion is to ensure that the CME propagation is modelled by DBM with enough confidence. It is important to highlight that the cone geometry of CME is not included in our calculation. We simply employ the 1-dimensional version of DBM in our calculation. Different versions of the DBM, that take into account the different cone geometries discussed in Dumbović et al. (2021), Schwenn (2006) can be seamlessly implemented in the calculation in the future by including additional free variables in the DBM inversion procedure. We have retrieved γ and w for 204 out of 213 ICMEs, which enables us to obtain robust statistics. The empirical PDF for the solar wind w is modelled using two separate distributions for slow and fast solar wind conditions respectively with a threshold value of $w = 500$ km/s for the fast solar wind. In Dumbović et al. (2018, 2021), Napoletano et al. (2018, 2022), a Gaussian distribution is assumed as input PDF for w. Here, we have used the threshold of $w = 500$ km/s for the fast solar wind speed, therefore, a normal distribution is no longer the ideal PDF. With this new threshold, the Student's t-distribution is the best choice for most

CME events. This latter finding is also supported by fitting PDFs for w in a single CME approach. In Figure 9, a histogram of the most suitable PDFs for w in individual CMEs approach is shown. Here, the Student's t-distribution is strongly biased by the fact of hard thresholding and the RSS values of Student's-t and normal distribution are fairly comparable, we therefore prefer the Gaussian PDF for the solar wind w. In our study, the mean value for slow solar wind speed is $w_{\text{slow}} = 370$ km/s with a standard deviation of 88 km/s, which is comparable with Napoletano et al. (2022). While for fast solar wind speed, the mean value is $w_{\text{fast}} = 579 \pm 68$ km/s and this value is somewhat higher than Napoletano et al. (2022). Notably, the median values are 386 km/s for the slow solar wind speed and 547 km/s for the fast solar wind scenario, respectively. It is important to realize that these mean values are marginally different from the median values because some extreme values are accepted in the inversion procedure. For instance, low values of some accepted solar wind speeds shift the mean leftward for the slow solar wind, whereas high values cause the mean to move rightward for the fast solar wind speed. The comparison of values of solar wind speed w for different previous studies is tabulated in Table 2. Also, the most probable value of solar wind speed obtained in different situations
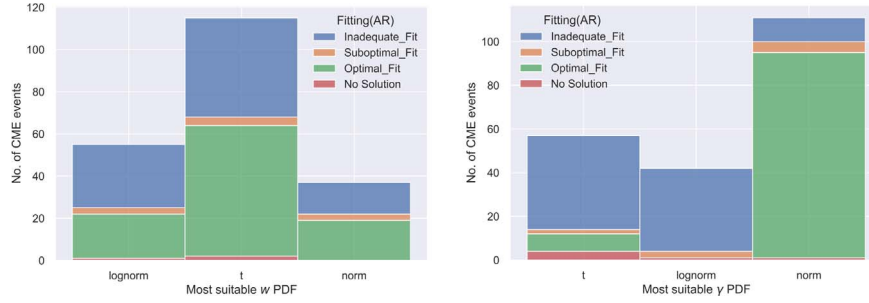
Figure 9. Histogram illustrating most suitable PDF for solar wind speed w and drag parameter γ in single CME approach. Within each PDF, various types of CME events are stratified and effectively stacked on top of one another.

Table 4. Statistical values for all CME events flagged as "Optimal Fit" analyzed together to obtain drag parameter γ and ambient solar wind speed w using DBM inversion procedure.

	Slow	Fast	Decelerating	Accelerating
Solar wind speed w (km/s)				
Mean	371	579	409	503
Median	386	547	402	494
σ	89	68	118	56
Drag parameter γ ($\times 10^{-7}$ km^{-1})				
Mean	0.86	1.26	0.93	1.59
Median	0.56	1.10	0.70	1.50
σ	0.74	0.73	0.76	0.58

like slow, fast, decelerating and accelerating is provided in Table 4.

The PDF for γ is up for discussion from the previous works of Čalogović et al. (2021), Dumbović et al. (2018, 2021) and Napoletano et al. (2018, 2022). One group employs a lognormal function, while the other group uses a Gaussian Function as input PDF. We have tried to fit the PDF on the entire dataset and single CME events, and our study has provided light on the preference for these two different functions. From Table 3, it is clear that lognormal distribution is the most favourable PDF as the RSS value is lower among other PDFs. On the contrary, when searching for the most suitable PDF in the single CME approach, the Gaussian PDF seems to be the best. In Figure 9, a histogram of the most suitable PDFs for γ in individual CME approaches is shown. A possible reason behind this discrepancy is the extensive dataset. Here, we have used a vast dataset of CMEs, which covers different ranges of mass and cross-sections of CME, also including almost two solar cycles' length of CME events resulting in several kinds of solar wind density fluctuations in CME propagation. The inclusion of all these background parameters in fitting a PDF through a dataset leads to the long-tailed lognormal function since the γ-parameter is a quantitative measure of the drag efficiency that depends on many factors such as the mass and the cross-section of the CME, and on the solar wind density (Vršnak et al., 2013). The values of drag parameters obtained in this study are quite higher compared to Čalogović et al. (2021), Dumbović et al. (2018, 2021). These higher values are the result of a long-tailed lognormal function. The statistical measurements for the values of drag parameter γ are tabulated in Table 4. It is important to note that one can not use the values mentioned in Table 4

directly as model input to predict CME arrival time and speed. To determine the CME arrival time and speed using DBM one has to use values provided in Table 3 using a transformation described in Appendix B.

The refined dataset and the updated method presented in this work allowed us to explore a larger part of the w–γ parameter space of the P-DBM model, including extreme values. We have investigated the possibility of γ being a function of the ICME kinematic properties (i.e., accelerating or decelerating) or the solar wind properties (i.e., fast or slow). While there seems to be some difference between accelerating or decelerating ICMEs (see Table 3 and Fig. 7), the statistics need to be more robust to draw strong conclusions. Therefore, similar efforts have been carried out for solar wind speed. The note-worthy iteration over a work of Napoletano et al. (2022) is to redetermine the solar wind type associated with a CME propagation in the heliosphere. Using 500 km/s as a threshold for the fast solar wind speed, we were able to infer distinguishable PDFs for fast and slow solar wind speed.

We suggest that the space weather community will benefit from our findings, especially the improved list of CME–ICME since it will provide a test bench to compare how well we can predict CME arrival time and impact. Also, the information associated with every CME ICME entry can help improve the accuracy and precision of other CME propagation models, by including other relevant parameters. For example, we want to use the Markov Chain Monte Carlo (MCMC) method in the future to constrain the PDF for w and γ. This catalogue's new entries are expected to play a relevant part in this work, promoting the convergence of Markov chains and boosting the performance of our strategy.

Data availability statement

The ICME catalogue built as a part of this work along with data visualisation and PDF analysis modules for implementing the DBM inversion procedure can be downloaded from https://zenodo.org/record/8063404 (Mugatwala et al., 2023).

References

Aquino M, Sreeja V. 2013. Correlation of scintillation occurrence with interplanetary magnetic field reversals and impact on Global Navigation Satellite System receiver tracking performance. *Space Weather* **11(5)**: 219–224. https://doi.org/10.1002/swe.20047.

Barbieri L, Mahmot R. 2004. October–November 2003's space weather and operations lessons learned. *Space Weather* **2(9)**: S09002. https://doi.org/10.1029/2004SW000064.

Bobra MG, Ilonidis S. 2016. Predicting coronal mass ejections using machine learning methods. *Astrophys J* **821(2)**: 127. https://doi.org/10.3847/0004-637X/821/2/127.

Brueckner G, Howard R, Koomen M, Korendyke C, Michels D, et al. 1995. The large angle spectroscopic coronagraph (LASCO) visible light coronal imaging and spectroscopy. *Sol Phys* **162**: 357–402. https://doi.org/10.1007/BF00733434.

Čalogović J, Dumbović M, Sudar D, Vršnak B, Martinić K, Temmer M, Veronig AM. 2021. Probabilistic Drag-Based Ensemble Model (DBEM) evaluation for heliospheric propagation of CMEs. *Sol Phys* **296(7)**: 114. https://doi.org/10.1007/s11207-021-01859-5.

Camporeale E. 2019. The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather* **17(8)**: 1166–1207. https://doi.org/10.1029/2018SW002061.

Cargill PJ. 2004. On the aerodynamic drag force acting on interplanetary coronal mass ejections. *Sol. Phys.* **221(1)**: 135–149. https://doi.org/10.1023/b:sola.0000033366.10725.a2.

Chen P. 2011. Coronal mass ejections: models and their observational basis. *Living Rev Sol Phys* **8**: 1–92. https://doi.org/10.12942/lrsp-2011-1.

Del Moro D, Napoletano G, Forte R, Giovannelli L, Pietropaolo E, Berrilli F. 2019. Forecasting the 2018 February 12th CME propagation with the P-DBM model: A fast warning procedure. *Ann Geophys* **61**: GM456. https://doi.org/10.4401/ag-7750.

Domingo V, Fleck B, Poland AI. 1995. The SOHO mission: An overview. *Sol Phys* **162**: 1–37. https://doi.org/10.1007/BF00733425.

Dumbović M, Čalogović J, Vršnak B, Temmer M, Mays ML, Veronig A, Piantschitsch I. 2018. The Drag-based Ensemble Model (DBEM) for coronal mass ejection propagation. *Astrophys J* **854(2)**: 180. https://doi.org/10.3847/1538-4357/aaaa66.

Dumbović M, Čalogović J, Martinić K, Vršnak B, Sudar D, Temmer M, Veronig A. 2021. Drag-based model (DBM) tools for forecast of coronal

mass ejection arrival time and speed. *Front Astron Space Sci* **8(639)**: 986. https://doi.org/10.3389/fspas.2021.639986.

Eyles C, Harrison R, Davis CJ, Waltham N, Shaughnessy B, et al. 2009. The heliospheric imagers onboard the STEREO mission. *Sol Phys* **254(2)**: 387–445. https://doi.org/10.1007/s11207-008-9299-0.

Gopalswamy N, Lara A, Lepping RP, Kaiser ML, Berdichevsky D, St. Cyr OC. 2000. Interplanetary acceleration of coronal mass ejections. *Geophys Res Lett* **27(2)**: 145–148. https://doi.org/10.1029/1999GL003639.

Gopalswamy N. 2009. Coronal mass ejections and space weather. In *Climate and Weather of the Sun-Earth System (CAWSES): Selected Papers from the 2007 Kyoto Symposium*. Terrapub, Tokyo, Japan, pp. 77–120. http://www.terrapub.co.jp/onlineproceedings/ste/CAWSES2007/index.html.

Gosling J, McComas D, Phillips J, Bame S. 1991. Geomagnetic activity associated with Earth passage of interplanetary shock disturbances and coronal mass ejections. *J Geophys Res Space Phys* **96(A5)**: 7831–7839. https://doi.org/10.1029/91JA00316.

Howard RA, Moses JD, Vourlidas A, Newmark JS, Socker DG, et al. 2008. Sun earth connection coronal and heliospheric investigation (SECCHI). *Space Sci Rev* **136(1–4)**: 67–115. https://doi.org/10.1007/s11214-008-9341-4.

Ivezić Ž, Connolly AJ, VanderPlas JT, Gray A. 2014. *Statistics, data mining, and machine learning in astronomy*, Princeton University Press. https://doi.org/10.1515/97814008489110.

Kaiser ML, Kucera TA, Davila JM, St OC, Cyr M Guhathakurta, Christian E. 2008. The STEREO mission: An introduction. *Space Sci Rev* **136**: 5–16. https://doi.org/10.1007/s11214-007-9277-0.

Koskinen H, Huttunen K. 2006. Geoeffectivity of coronal mass ejections. *Space Sci Rev* **124**: 169–181. https://doi.org/10.1007/s11214-006-9103-0.

Koskinen H, Huttunen K. 2007. Geoeffectivity of coronal mass ejections. In *Solar dynamics and its effects on the heliosphere and earth*, pp. 169–181. https://doi.org/10.1007/s11214-006-9103-0.

Liu Y, Thernisien A, Luhmann JG, Vourlidas A, Davies JA, Lin RP, Bale SD. 2010. Reconstructing coronal mass ejections with coordinated imaging and in situ observations: Global structure, kinematics, and implications for space weather forecasting. *Astrophys J* **722(2)**: 1762. https://doi.org/10.1088/0004-637X/722/2/1762.

Liu J, Ye Y, Shen C, Wang Y, Erdélyi R. 2018. A new tool for CME arrival time prediction using machine learning algorithms: CAT-PUMA. *Astrophys J* **855(2)**: 109. https://doi.org/10.3847/1538-4357/aaae69.

Manchester W, Kilpua EK, Liu YD, Lugaz N, Riley P, Török T, Vršnak B. 2017. The physical processes of CME/ICME evolution. *Space Sci Rev* **212**: 1159–1219. https://doi.org/10.1007/s11214-017-0394-0.

Manoharan P. 2006. Evolution of coronal mass ejections in the inner heliosphere: A study using white-light and scintillation images. *Sol Phys* **235**: 345–368. https://doi.org/10.1007/s11207-006-0100-y.

Mugatwala R, Chierichini S, Francisco G, Napoletano G, Foldes R, Giovannelli L, Gasperis GD, Camporeale E, Erdélyi R, Moro DD. 2023. *Wolpes11/PDBM-project-for-ICMEs*. https://doi.org/10.5281/zenodo. 8063404.

Napoletano G, Forte R, Moro DD, Pietropaolo E, Giovannelli L, Berrilli F. 2018. A probabilistic approach to the drag-based model. *J Space Weather Space Clim* **8**: A11. https://doi.org/10.1051/swsc/2018003.

Napoletano G, Foldes R, Camporeale E, de Gasperis G, Giovannelli L, Paouris E, Pietropaolo E, Teunissen J, Tiwari AK, Moro DD. 2022. Parameter distributions for the drag-based modeling of CME propagation. *Space Weather* **20(9)**: e2021SW002925. https://doi.org/10.1029/2021sw002925.

Odstrcil D, Vandas M, Pizzo VJ, MacNeice P. 2003. Numerical simulation of interacting magnetic flux ropes. *AIP Conf Proc* **679**: 699–702. https://doi.org/10.1063/1.1618690.

Paouris E, Mavromichalaki H. 2017. Effective acceleration model for the arrival time of interplanetary shocks driven by coronal mass ejections. *Sol Phys* **292**: 1–11. https://doi.org/10.1007/s11207-017-1212-2.

Paouris E, Čalogovic J, Dumbović M, Mays ML, Vourlidas A, Papaioannou A, Anastasiadis A, Balasis G. 2021. Propagating conditions and the time of ICME arrival: A comparison of the effective acceleration model with ENLIL and DBEM models. *Sol Phys* **296(1)**: 12. https://doi.org/10.1007/s11207-020-01747-4.

Papaioannou A, Sandberg I, Anastasiadis A, Kouloumvakos A, Georgoulis MK, Tziotziou K, Tsiropoula G, Jiggens P, Hilgers A. 2016. Solar flares, coronal mass ejections and solar energetic particle event characteristics. *J Space Weather Space Clim* **6**: A42. https://doi.org/10.1051/swsc/2016035.

Piersanti M, Alberti T, Bemporad A, Berrilli F, Bruno R, et al. 2017. Comprehensive analysis of the geoeffective solar event of 21 June 2015: Effects on the magnetosphere, plasmasphere, and ionosphere systems. *Sol Phys* **292**: 1–56. https://doi.org/10.1007/s11207-017-1186-0.

Pomoell J, Poedts S. 2018. EUHFORIA: European heliospheric forecasting information asset. *J Space Weather Space Climate* **8**: A35. https://doi.org/10.1051/swsc/2018020.

Pulkkinen T. 2007. Space weather: terrestrial perspective. *Living Rev Sol Phys* **4**: 1–60. https://doi.org/10.12942/lrsp-2007-1.

Richardson IG, Cane HV. 2010. Near-earth interplanetary coronal mass ejections during solar cycle 23 (1996–2009): Catalog andsummary of properties. *Sol Phys* **264(1)**: 189–237. https://doi.org/10.1007/s11207-010-9568-6.

Riley P, Mays ML, Andries J, Amerstorfer T, Biesecker D, et al. 2018. Forecasting the arrival time of coronal mass ejections: analysis of the CCMC CME scoreboard. *Space Weather* **16(9)**: 1245–1260. https://doi.org/10.1029/2018sw001962.

Rollett T, Möstl C, Isavnin A, Davies JA, Kubicka M, Amerstorfer UV, Harrison RA. 2016. ElEvoHI: A novel CME prediction tool for heliospheric imaging combining an elliptical front with drag-based model fitting. *Astrophys J* **824(2)**: 131. https://doi.org/10.3847/0004-637X/824/2/13.

Sachdeva N, Subramanian P, Colaninno R, Vourlidas A. 2015. CME propagation: Where does aerodynamic drag "take over"? *Astrophys J* **809(2)**: 158. https://doi.org/10.1088/0004-637x/809/2/158.

Sachdeva N, Subramanian P, Vourlidas A, Bothmer V. 2017. CME dynamics using STEREO and LASCO observations: The relative importance of Lorentz forces and solar wind drag. *Sol Phys* **292(9)**: 118. https://doi.org/10.1007/s11207-017-1137-9.

Schrijver CJ, Siscoe GL. 2010. *Heliophysics: Space storms and radiation: causes and effects*. Cambridge University Press.

Schwenn R. 2006. Space weather: The solar perspective. *Living Rev Sol Phys* **3**: 1–72. https://doi.org/10.12942/lrsp-2006-2.

Shea M, Smart D. 1998. Space weather: The effects on operations in space. *Adv Space Res* **22(1)**: 29–38. https://doi.org/10.1016/S0273-1177(97)01097-1.

Sreeja V. 2016. Impact and mitigation of space weather effects on GNSS receiver performance. *Geosci Lett* **3(1)**: 24. https://doi.org/10.1186/s40562-016-0057-0.

Taskesen E. 2023. *Distfit is a python library for probability density fitting*. https://doi.org/10.5281/zenodo.7650685.

Temmer M. 2021. Space weather: The solar perspective. *Living Rev Sol Phys* **18(1)**: 4. https://doi.org/10.1007/s41116-021-00030-3.

Tsurutani BT, Gonzalez WD, Tang F, Akasofu SI, Smith EJ. 1988. Origin of interplanetary southward magnetic fields responsible for major magnetic storms near solar maximum (1978–1979). *J Geophys Res Space Phys* **93(A8)**: 8519–8531. https://doi.org/10.1029/JA093iA08p08519.

VanderPlas J, Connolly AJ, Ivezić Ž, Gray A. 2012. Introduction to astroML: Machine learning for astrophysics. In *2012 conference on intelligent data understanding*. IEEE, pp. 47–54. https://doi.org/10.1109/CIDU.2012.6382200.

Veettil SV, Cesaroni C, Aquino M, De Franceschi G, Berrili F, et al. 2019. The ionosphere prediction service prototype for GNSS users. *J Space Weather Space Clim* **9**: A41. https://doi.org/10.1051/swsc/2019038.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, et al. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods* **17(3)**: 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vourlidas A, Patsourakos S, Savani NP. 2019. Predicting the geoeffective properties of coronal mass ejections: current status, open issues and path forward. *Philos Trans Royal Soc A: Math Phys Eng Sci* **377(2148)**: 20180096. https://doi.org/10.1098/rsta.2018.0096.

Vršnak B. 2001. Dynamics of solar coronal eruptions. *J Geophys Res Space Phys* **106(A11)**: 25249–25259.

Vršnak B, Ruždjak D, Sudar D, Gopalswamy N. 2004. Kinematics of coronal mass ejections between 2 and 30 solar radii-What can be learned about forces governing the eruption? *A&A* **423(2)**: 717–728.

Vršnak B, Žic T, Vrbanec D, Temmer M, Rollett T, et al. 2013. Propagation of interplanetary coronal mass ejections: The drag-based model. *Sol Phys* **285(1–2)**: 295–315.

Wang P, Zhang Y, Feng L, Yuan H, Gan Y, Li S, Lu L, Ying B, Gan W, Li H. 2019. A new automatic tool for CME detection and tracking with machine-learning techniques. *Astrophys J Suppl Ser* **244(1)**: 9. https://doi.org/10.3847/1538-4365/ab340c.

Webb DF, Howard TA. 2012. Coronal mass ejections: Observations. *Living Rev Sol Phys* **9(1)**: 1–83. https://doi.org/10.12942/lrsp-2012-3.

Wu C-C, Fry C, Wu S, Dryer M, Liou K. 2007. Three-dimensional global simulation of interplanetary coronal mass ejection propagation from the Sun to the heliosphere: Solar event of 12 May 1997. *J Geophys Res Space Phys* **112(A9)**: A09104. https://doi.org/10.1029/2006JA012211.

Appendix A

Description of revised dataset

As mentioned above, the DBM inversion procedure requires initial position r_0, target position r_{1AU}, transit time t_{1AU}, initial speed v_0 and arrival speed v_a to obtain w and γ. For the purpose of this work, we have used the CME–ICME dataset from Napoletano et al. (2022). This dataset contains all the required input quantities for the DBM inversion procedure. This dataset consists of 213 CME–ICME pairs from the year 1997 to 2018, which cover a time span of two solar cycles 23 and 24. In this dataset, information about the kinematic properties of CMEs at launch time was retrieved from the SOHO/LASCO CME Catalog[1]. While arrival time and speed of the related ICMEs have been obtained from the Richardson & Cane (2010).

As mentioned in Section 2.2, the uncertainties associated with different quantities are included in the inversion procedure. SOHO/LASCO catalogue provides CME speed in the plane of sky (POS) but to make a DBM forecast more accurate de projected speed has been used in the calculation. De projected radial speed has been obtained using equation (1) of Gopalswamy (2009). A more detailed explanation is given in Appendix A2 of Napoletano et al. (2022). Associated solar wind speed type (column: SW_type) for each event is hypothesized by determining the presence of a coronal hole close to the CME source region (see Appendix A3 of Napoletano et al., 2022). The description of different columns in the dataset and their source work is provided in a Table A.1.

Appendix B

Mathematical description of lognormal distribution

To find the parameters of lognormal PDF we have used a Python package named *distfit* Taskesen (2023) which relies on *SciPy* Virtanen et al. (2020). The standardized form of the lognormal function is given as:

$$f(x, s) = \frac{1}{sx\sqrt{2\pi}} \exp\left(\frac{-\ln^2 x}{2s^2}\right). \quad (B.1)$$

To shift and/or scale the above distribution function, SciPy or distfit use two more input parameters namely *loc* and *scale*. With these 2 more parameters, the new function will be:

$$f(x, s, \text{loc}, \text{scale}) = \frac{f(y, s)}{\text{scale}}, \quad (B.2)$$

where $y = \dfrac{x - \text{loc}}{\text{scale}}$. Suppose, a variable X is following a normal distribution with parameters μ and σ. Then, lognormally

---

1 https://cdaw.gsfc.nasa.gov/CME_list/.

Table A.1. Column description of the ICME dataset created as a part of this work.

Name	Keyword	Description	Source
LASCO start	LASCO_Start	First CME appearance in LASCO C2/C3 coronagraphs	LASCO/CDAW[1]
Start date	Start_Date	Time when CME reaches to 20 R$_\odot$	Napoletano et al. (2022)
Arrival date	Arrival_Date	Estimated arrival time of ICME using insitu signatures	R & C[2]
Plasma event duration	PE_duration	End of ICME plasma signatures after col 3 is recorded	R & C
Arrival speed	Arrival_v	ICME arrival speed at L1 (km/s)	R & C
Transit time	Transit_time	(h) Computed between col 1 and col 3	Napoletano et al. (2022)
Transit time error	Transit_time_err	(h) Error associated to the start date of CME	Napoletano et al. (2022)
LASCO date	LASCO_Date	Most likely associated CME observed by LASCO	LASCO/CDAW
LASCO speed	LASCO_v	(km/s) Speed correspond to the fastest moving point of CME in LASCO FOV	LASCO/CDAW
Position angle	LASCO_pa	(deg) Counterclockwise (from solar North) angle of appearance into coronographs	LASCO/CDAW
Angular width	LASCO_da	(deg.) Angular expansion of CME into coronographs	LASCO/CDAW
Halo	LASCO_halo	If LASCO_da is >270° then "FH" (full halo), if >180° "HH" (half halo), if >90° "PH"(partial halo), otherwise "NO"	LASCO/CDAW
De-projected speed	v_r	(km/s) De-projected CME speed	Napoletano et al. (2022)
De-projected speed error	v_r_err	(km/s) Uncertainty of CME initial speed	Napoletano et al. (2022)
Theta source	Theta_source	(arcsec) Longitude of the most likely source of CME	Napoletano et al. (2022)
Phi source	Phi_source	(arcsec) Co-latitude of the most likely source of CME	Napoletano et al. (2022)
Source POS error	source_err	(deg.) Uncertainty of the most likely CME source	Napoletano et al. (2022)
POS source angle	POS_source_angle	(deg.) Principal angle of the most likely CME source	Napoletano et al. (2022)
Relative width	rel_wid	(rad.) De-projected width of CME	Napoletano et al. (2022)
Mass	Mass	(gm) Estimated CME Mass	LASCO/CDAW
Solar wind type (CH)	SW_type	Solar wind (slow, S, or fast, F) interacting with the ICME based on the presence of coronal hole near CME location	Napoletano et al. (2022)
Bz	Bz	(nT) z-component of magnetic field at L1 and CME arrival time	R & C
Dst	DST	Geomagnetic Dst index recorded at CME arrival	R & C
Statistical de-projected speed	v_r_stat	(km/s) Statistical de-projected CME speed, that is, $v\_r\_stat = LASCO\_v \times 1.027 + 41.5$	
Acceleration	Accel.	(m/s^2) Residual acceleration at last CME observation	Napoletano et al. (2022)
Analytical wind	Analyitic_w	(km/s) Solar wind from DBM exact inversion	Napoletano et al. (2022)
Analytical gamma	Analyitic_gamma	(km^{-1}) Drag parameter, γ, from DBM exact inversion	Napoletano et al. (2022)
Transit Time (Simulated)	T1_Sim	(h) Transit time calculated using P-DBM	This Work
Transit Time error (Simulated)	T1_Sim_err	(h) Error associated with transit time in P-DBM	This Work
Impact Speed (Simulated)	V1_Sim	(km/s) Calculated CME arrival speed using P-DBM	This Work
Impact Speed error (Simulated)	V1_Sim_err	(km/s) Error associated with arrival speed in P-DBM	This Work
Solar Wind Speed	W_Sim	(km/s) Mean value of solar wind speed from inversion procedure	This Work
Solar Wind Speed Error	W_Sim_err	(km/s) Standard deviation of solar wind speed from inversion procedure	This Work
Gamma Simulated	Gamma_Sim_s	(km^{-1}) "s" parameter for lognormal PDF	This Work
Gamma Error Simulated	Gamma_Sim_loc	(km^{-1}) "loc" parameter for lognormal PDF	This Work
Gamma Simulated (log)	Gamma_Sim_scale	"scale" parameter for lognormal PDF	This Work
Optimal Transit Time	T1_opt	Minimally deviated transit time compared to observed one	This Work
Optimal Impact Speed	V1_opt	V1 correspond to T1_opt	This Work
Optimal W	W_opt	W correspond to T1_opt	This Work
Optimal gamma	Gamma_opt	Gamma correspond to T1_opt	This Work

(Continued on next page)

Table A.1. (Continued)

Name	Keyword	Description	Source
Optimal V_r	V_r_opt	V_r correspond to T1_opt	This Work
W CI min	W99_min	Minimum value of 99% confidence interval for w	This Work
W CI max	W99_max	Maximum value of 99% confidence interval for w	This Work
Gamma CI min	Gamma99_min	Minimum value of 99% confidence interval for gamma	This Work
Gamma CI max	Gamma99_max	Maximum value of 99% confidence interval for gamma	This Work
CME Type (V_r_opt)	CME_type	CME type based on W_sim (Accelerating/ Decelerating)	This Work
CME Type (V_r)	CME_type_v0	CME type based on W_opt (accelerating: A/decelerating D)	This Work
Solar wind Type (Wth)	Wind_type	Solar wind (based on threshold value) interacting with ICME	This Work
Target distance	R1(AU)	(AU) Sun-Earth Distance at CME start date (Col2)	This Work
Fitting	Fitting (AR)	Goodness of Inversion procedure: Optimal/ Suboptimal/Inadequate	This Work
Acceptance Rate	Acceptance_Rate	Acceptance rate of inversion procedure	This Work
Best W PDF	Best_fit_W	Most suitable PDF for W	This Work
Best gamma PDF	Best_fit_gamma	Most suitable PDF for gamma	This Work

[1] LASCO/CDAW: https://cdaw.gsfc.nasa.gov/CME_list/
[2] Richardson and Cane ICME list (R & C): https://izw1.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm

distributed variable $Y = \exp(X)$ has $\mu = \ln(\text{scale})$ and $\sigma = s$. The simplified version of formula (B.2) is given as follow:

$$f(x, s, \text{loc}, \text{scale}) = \frac{1}{s(x - \text{loc})\sqrt{2\pi}} \exp\left[-\left(\frac{(\ln(x - \text{loc}) - \mu)}{\sqrt{2}s}\right)^2\right],$$

(B.3)

while a lognormal function used by Napoletano et al. (2018) is

$$f(x,) = \frac{1}{s\sqrt{2\pi}} \exp\left[-\left(\frac{(\ln x - \mu)}{\sqrt{2}s}\right)^2\right].$$

(B.4)