









RESEARCH ARTICLE

Updating Measures of CME Arrival Time Errors

10.1029/2024SW003951

C. Kay^{1,2,3} , E. Palmerio⁴ , P. Riley⁴ , M. L. Mays² , T. Nieves-Chinchilla², M. Romano^{2,3}, Y. M. Collado-Vega² , C. Wiegand² , and A. Chulaki^{2,3}

Key Points:

- From 1702 arrival time predictions we find a bias of -2.5 hr, mean absolute error of 13.2 hr, and standard deviation of 17.4 hr
- The routinely-submitted models all perform fairly similar but there is more variation in models with smaller data sets
- We find evidence of late predictions for coronal mass ejections (CMEs) with short transit times, and early predictions for CMEs with long transit times

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. Kay,
Christina.Kay@jhuapl.edu

Citation:

Kay, C., Palmerio, E., Riley, P., Mays, M. L., Nieves-Chinchilla, T., Romano, M., et al. (2024). Updating measures of CME arrival time errors. *Space Weather*, 22, e2024SW003951. <https://doi.org/10.1029/2024SW003951>

Received 4 APR 2024

Accepted 3 JUL 2024

¹The Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA, ²Heliophysics Science Division, NASA Goddard Space Flight Center, Greenbelt, MD, USA, ³Department of Physics, The Catholic University of America, Washington, DC, USA, ⁴Predictive Science Inc., San Diego, CA, USA

Abstract Coronal mass ejections (CMEs) drive space weather effects at Earth and the heliosphere. Predicting their arrival is a major part of space weather forecasting. In 2013, the Community Coordinated Modeling Center started collecting predictions from the community, developing an Arrival Time Scoreboard (ATSB). Riley et al. (2018, <https://doi.org/10.1029/2018sw001962>) analyzed the first 5 years of the ATSB, finding a bias of a few hours and uncertainty of order 15 hr. These metrics have been routinely quoted since 2018, but have not been updated despite continued predictions. We revise analysis of the ATSB using a sample 3.5 times the size of that in the original study. We find generally the same overall metrics, a bias of -2.5 hr, mean absolute error of 13.2 hr, and standard deviation of 17.4 hr, with only a slight improvement comparing between the previously-used and new sets. The most well-established, frequently-submitted model results tend to outperform those from seldomly-contributed models. These “best” models show a slight improvement over the 11 year span, with more scatter between the models during early times and a convergence toward the same error metrics in recent years. We find little evidence of any correlations between the arrival time errors and any other properties. The one noticeable exception is a tendency for late predictions for short transit times and vice versa. We propose that any model-driven systematic errors may be washed out by the uncertainties in CME reconstructions in characterization of the background solar wind, and suggest that improving these may be the key to better predictions.

Plain Language Summary Coronal mass ejections (CMEs) are huge explosions that erupt from the Sun and travel through the solar system. It is important to have warning of when they reach Earth, so many researchers model the arrival time (AT) of CMEs. Many predictions are collected in the AT Scoreboard (ATSB), and the general errors from the first 5 years of the ATSB were determined in 2018. These errors are very important as a measure of the uncertainty of predictions, but they have not been updated in over 6 years. Now the data set is over 3.5 times larger so we determine updated results but ultimately only see a small change in the errors. We find a bias of only a few hours, so no strong tendency for either early or late predictions. We find an average absolute error of 13.2 hr but a wide range about this value for individual cases. We do see some evidence that the most commonly used models have improved over time, but the change is very small and there are many factors affecting whether it is significant or not. We suggest that improving how we reconstruct CME properties and model the background solar wind may improve predictions.

1. Introduction

Coronal mass ejections (CMEs) from the Sun can be listed among the major drivers of geomagnetic effects at Earth. As such a large part of space weather research focuses on predicting their propagation through the heliosphere and/or their structure upon impact (e.g., Kilpua et al., 2019; Vourlidis et al., 2019). Parameters that are of particular interest for real-time predictions include whether a CME will impact at all (also known as *the hit/miss problem*), the time and speed of arrival (also known as *the arrival time problem*), as well as the magnitude and duration of southward magnetic fields (also known as *the B_z problem*), all known to contribute prominently to the timing and intensity of a geomagnetic storm (e.g., Gonzalez et al., 1994; Kilpua et al., 2017). Despite recent progress in modeling CME magnetic fields in situ in the form of hindcasts in real-time-like settings (e.g., Asvestari et al., 2021; Kay et al., 2017; Maharana et al., 2023; Palmerio et al., 2021, 2023; Scolini et al., 2019, 2020), predicting B_z remains one of the major challenges in space weather research, while most efforts continue to target the hit/miss and arrival time (AT) problems.

© 2024. The Author(s).

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

The overall procedure behind the generation of an AT prediction is relatively straightforward: (a) obtain a certain set of input parameters from observational data, (b) forward model the propagation of the CME via empirical or physics-based relations (or a combination of the two), and (c) produce an estimate of the impact at the desired target, often by means of synthetic observables. This sequence of steps toward a prediction can be achieved via a large variety of models, which may differ by the methodologies employed and/or by the complexity of the physics under assumption. For example, X. Zhao and Dryer (2014) categorized the wide range of existing codes as empirical models (e.g., Gopalswamy et al., 2001; Kim et al., 2007; Vandas et al., 1996), expansion speed models (which are nevertheless a variant of empirical ones; Schwenn et al., 2005), drag-based models (e.g., Rollett et al., 2016; Shi et al., 2015; Vršnak et al., 2013), physics-based models (e.g., Corona-Romero et al., 2017; Hess & Zhang, 2015; Paouris & Vourlidas, 2022), and magnetohydrodynamic (MHD) models (e.g., Mayank et al., 2024; Pomoell & Poedts, 2018; Riley & Ben-Nun, 2022). In addition, more recent developments include machine-learning (ML) models (e.g., Alobaid et al., 2022; Liu et al., 2018; Y. Wang et al., 2019), and it is worth remarking that there exists a class of models that may be considered “hybrid,” as they bring together aspects from multiple categories (e.g., Barnard & Owens, 2022; Kay et al., 2022; Singh et al., 2023).

Since models that are mostly oriented toward predicting the AT of CMEs tend to be computationally efficient and can be rapidly run even in ensemble mode (e.g., Amerstorfer et al., 2018; Dumbović et al., 2018; Mays et al., 2015), they can be tested and validated in real-time applications. Since 2013, the Community Coordinated Modeling Center (CCMC) at NASA's Goddard Space Flight Center (GSFC) maintains a CME Scoreboard (viewable at <https://kauai.ccmc.gsfc.nasa.gov/CMEScoreboard>) that catalogs predictions from a collection of international researchers employing a variety of CME AT models. We will refer to this as the AT Scoreboard (hereafter ATSB) to differentiate from other scoreboards available through the CCMC. After registering their models, users can submit predictions for any observed CME that they choose to simulate. Predictions are collected until the time of CME arrival and the collected results are used to determine the “Average of all Methods” value. Not all submitted predictions end up corresponding to an observed CME arrival, as there are numerous non-impact cases or false alarms.

One of the most common combination of models used for predictions is the WSA–Enlil + Cone (hereafter WEC) architecture—employed by space weather agencies such as in the USA, UK, Korea, and Australia (e.g., Pizzo et al., 2011). It consists of two spatial domains, that is, the semi-empirical Wang-Sheeley-Arge (WSA; e.g., Arge & Pizzo, 2000; Arge et al., 2004) coronal module that covers the range 1–21.5 R_{\odot} and the MHD Enlil (e.g., Odstrcil, 2003; Odstrcil et al., 2004) heliospheric module that simulates the heliosphere from 21.5 R_{\odot} or 0.1 au outwards, as well as the so-called cone model (e.g., X. P. Zhao et al., 2002; Xie et al., 2004) description of CMEs as hydrodynamic pulses. CMEs are inserted at the interface layer between WSA and Enlil, and their non-magnetized nature allows for relatively expeditious predictions of their AT and speed at one or more locations of interest (see also Odstrcil, 2023), despite not being able to accurately predict their internal magnetic fields. Wold et al. (2018, hereafter W18) analyzed predictions that the CCMC made using the WEC model for impacts expected at Earth or either of the STEREO satellites between March 2010 and December 2016, a total of 273 events with corresponding CMEs observed in situ. They found a mean error of -4.0 hr, suggesting a bias toward early predictions, and a mean absolute error (MAE) of 10.4 hr.

Riley et al. (2018, hereafter R18) built upon the work of W18, analyzing all of the predictions submitted by the community in the first 6 years of the ATSB (up to 11 May 2018), not just those from the CCMC itself. Combining results from 32 sources, R18 found mean errors between -25 and 10 hr and MAEs between 1.5 and 25 hr. Many of the largest errors correspond to models with a limited number of submitted predictions, while restricting the statistics to the six models with the most submissions yields a mean error of -2.8 hr and a MAE of 14.0 hr, similar to the results of W18. This suggests one should expect an average uncertainty of order 10 hr from a well-established model, as the standard deviation in the errors tends to be 15–20 hr, indicating a wide range in the spread of actual results about the average value. R18 also found no evidence for any improvement in the predictions between 2013 and 2018.

R18 has become one of the most frequently cited sources for the uncertainty and errors in AT predictions, but the metrics have not been updated with the ensuing half-decade of new data generated by the community. This work builds upon the analysis of R18 by including predictions submitted to the ATSB up until the end of 2023, thus adding an additional 5.5 years of results to the existing statistics. We describe the data set in Section 2 and compare our results with those of R18 in Section 3. In the remaining sections we look at the variations in the errors

over time (Section 4), with CME properties (Section 5), and by model (Section 6). Our discussions and conclusions are drawn in Sections 7 and 8, respectively.

2. Data Set

2.1. Data Collection

The ATSB lists all of the submitted predictions grouped by event, which is identified using the time of the corresponding observed coronal CME. The ATSB event header lists the “Actual Shock AT” for events with an observed impact, or states “This CME was not detected at Earth!” for non-impacting events. The ATSB defines the observed CME arrival as the CME-driven shock signature. When a shock is not observed, the arrival is defined as the observed discontinuity or magnetic cloud signature. Underneath each header is the list of all predictions submitted for that case, as well as an “Average of all Methods” combining the results of all the other predictions.

At a minimum, each prediction includes the predicted AT, the time of submission, the method (name of the model), and who submitted the prediction. There are also the options to include error bars on the AT, the confidence, and/or the predicted geomagnetic storm parameters (range in the expected maximum K_p index). The confidence in the prediction is expressed as a percentage, which represents the likelihood that the CME will actually arrive, not the confidence in the prediction of the timing. The specific determination of the confidence varies between models, often for ensembles it is the percentage of impacting cases out of the total number of ensemble members.

The predictions are separated into individual pages based on calendar year. We downloaded the HTML source code for each page and wrote a Python script to process the results into a CSV file. This data is archived via Zenodo, with the link available in the Data Availability Statement section. We find a total of 2,617 predictions from 39 different models between 15 May 2013 and 31 December 2023. We pick the end of 2023 as a “convenient” albeit somewhat arbitrary cutoff and suggest that results should be continually updated and documented online on at least a yearly basis. We note that the CCMC has now created an Application Programming Interface (API) that can be used to directly query the database. We describe the use of the API in Supporting Information S1.

2.2. Individual Models

Table 1 lists the 39 included models/users, along with the time range each catalog covers and the number of predictions n_p that were submitted using that model. We note that we will refer to these individual sets as “models” but the labels refer to both models and/or the group that submitted the results. The ATSB typically uses the naming convention of “Model (Group)” or just “Model” in the case of models used by a single group. Within the ATSB, some results are listed as “Other (Group)” if they do not use a named model, here we refer to those simply as “Group.” More details on the individual models can be found at <https://ccmc.gsfc.nasa.gov/scoreboards/cme> or in Section 2.1 of R18. We list the reference for each model in Table 1 where appropriate, some of the agency predictions do not have a dedicated publication, particularly those that rely on WEC. We use the ID number on the left within Table 1 to identify different models in several figures in Section 6.

As done in the ATSB, we distinguish between predictions made by the NASA GSFC Space Weather Research Center (SWRC) and the Moon to Mars (M2M) Space Weather Analysis Office. When the ATSB was first created, SWRC led the AT predictions at NASA GSFC. In 2020 the M2M Space Weather Analysis Office was established at NASA GSFC. All real-time analysis activities carried out by the SWRC as part of the CCMC have been transitioned to the M2M Space Weather Analysis Office and the ATSB switched to using M2M Space Weather Analysis Office predictions in January 2021. This office employs space weather analysts with more formalized training, as compared to the diverse team of students and scientists with varying levels of training/experience who supported experimental forecasting activities within the SWRC. The M2M Space Weather Analysis Office is responsible for closing out the CMEs on the scoreboard by identifying the observed arrival times of the CMEs (although any registered power user may also do so). For complex arrival signatures, the M2M Space Weather Analysis Office will regularly consult with CME arrival signature experts based at NASA GSFC within the Large Scale Structures Originating from the Sun research group. All ATSB entries are subject to both uncertainties in determining the observed CME AT from data, and determining the predicted arrival times from the model/method

Table 1
List of Individual Models Contributing to the Arrival Time Scoreboard

	Name	Time range	n_p	Reference
1	Anemomilos	2013 March–2022 May	28	Tobiska et al. (2013)
2	Average of all Methods	2013 March–2023 December	474	–
3	BHV	2013 March–2014 January	4	Bothmer and Schwenn (1998)
4	British Geological Survey	2017 September	2	–
5	CAT-PUMA	2018 February–2021 November	8	Liu et al. (2018)
6	CMEFM v.0.1	2023 Jun–2023 December	27	–
7	COMESSEP	2014 January–2015 March	7	Crosby et al. (2012)
8	Cone + HAF (SEPC, NSSC, CAS)	2020 December–2023 December	32	J. Wang et al. (2018)
9	DBM	2013 March–2021 December	19	Vršnak et al. (2013)
10	DBM + ESWF	2016 May–2019 May	5	Rotter et al. (2015)
11	EAM (Effective Acceleration Model)	2017 May–2023 December	268	Paouris and Mavromichalaki (2017)
12	ELEvo	2015 Jun–2023 November	5	Möstl et al. (2015)
13	ELEvoHI	2020 January–2023 November	4	Rollett et al. (2016)
14	ESA	2013 March–2014 January	3	Gopalswamy et al. (2005)
15	Ensemble WEC (GSFC SWRC)	2013 September–2020 December	61	Zheng and Rastaetter (2015)
16	Ensemble WEC (NASA M2M)	2021 February–2023 December	145	–
17	Expansion Speed Prediction Model	2014 January–2014 July	4	Schwenn et al. (2005)
18	H3DMHD (HAFv.3 + 3DMHD)	2013 March	1	Wu et al. (2011)
19	HAFv2w	2013 April	1	Smith et al. (2009)
20	HUXt	2023 December	3	Barnard and Owens (2022)
21	ips.gov.au	2013 March–2015 November	3	–
22	IZMIRAN	2023 December	2	–
23	NSSC SEPC	2017 September–2018 February	3	–
24	Ooty IPS	2017 September	2	–
25	Other	2017 November–2018 July	3	–
26	Rice-ENLIL Dst	2014 June	1	–
27	SAO Crowdsourcing	2014 April–2014 September	3	–
28	SARM	2014 February–2023 December	109	Núñez et al. (2016)
29	SIDC	2013 March–2023 December	271	–
30	SPM	2015 November–2023 July	55	Feng and Zhao (2006)
31	SPM2	2015 October–2023 July	57	X. H. Zhao and Feng (2014)
32	STOA	2013 April–2014 January	8	McKenna-Lawlor et al. (2006)
33	WEC	2013 September–2020 August	11	–
34	WEC (BoM)	2015 November–2023 November	91	–
35	WEC (GSFC SWRC)	2013 March–2020 December	138	Zheng and Rastaetter (2015)
36	WEC (KSWC)	2014 June–2021 March	34	–
37	WEC (Met Office)	2014 June–2023 December	236	–
38	WEC (NASA M2M)	2021 January–2023 December	323	–
39	WEC (NOAA/SWPC)	2013 March–2023 December	166	Pizzo et al. (2011)

outputs. We also note that the M2M Space Weather Analysis Office, and formerly the SWRC, submit predictions to the ATSB on the behalf of other organizations like SWPC and SIDC (which are listed under the name of the original source).

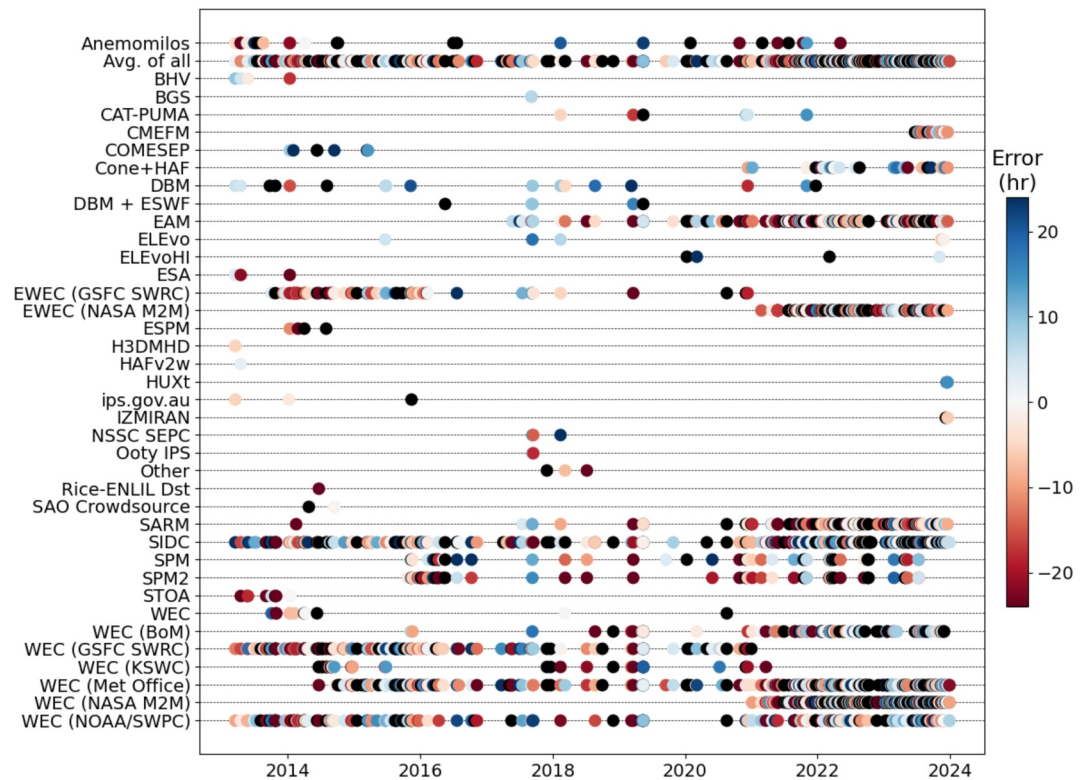


Figure 1. Temporal span of the individual models within the Arrival Time Scoreboard. Each dot represents an individual prediction and is colored according to the unsigned error in the arrival time. Predictions without a corresponding observed impact are shown in black.

We note that SWPC and other forecasting groups make a distinction between a prediction, such as direct model output, compared to a forecast that is produced by a human who may use model outputs together with data, previous experience, and any other information. The CME Scoreboard contains both predictions and forecasts, but for simplicity we are not distinguishing them in this evaluation. In particular, the UKMO enters their own CME AT submissions to the scoreboard and these are forecasts. SWPC entries are submitted by M2M Space Weather Analysis Office analysts from model output predictions taken from the SWPC website, and are not forecasts. Two additional caveats are that not all SWPC runs are posted to their website, or SWPC may post multiple runs for an event (typically the last is the “official” run) and not all of these may be captured by M2M Space Weather Analysis Office. Ideally, in the future, groups who produce both forecasts and predictions would provide both to the scoreboard (labeled), such that more detailed comparisons can be made.

We see a wide range in n_F , which is the total number of predictions (including both cases with and without observed impacts). There are 9 models with more than 100 predictions and 20 models with fewer than 10 predictions. Some models have predictions submitted over the full duration of this study whereas others have results only submitted for a single month. To better visualize the variety of the individual model data sets, Figure 1 shows a timeline of the predictions submitted for each model. Each dot represents an individual prediction and is colored by the error in the AT (predicted minus observed so positive is a late prediction). Non-impacting cases are shown in black.

Figure 1 highlights the variations between the data sets from different models. We see the highest density of points toward the bottom, which corresponds to all of the WEC models that are routinely submitted from different operational centers. No immediate patterns jump out in terms of trends in the AT error, but this figure can only give a broad overview and we will further investigate trends below.

3. Quantifying the Data Set

We begin our investigation by comparing the size of the new catalog as compared to R18. Table 2 provides a breakdown of the total number of events and the split between the original R18 set and the new set. The first three

Table 2
General Comparison of Data Sets

Set	n_P	$n_{P,I}$	$n_{P,NI}$	n_{CME}	$n_{CME,I}$	n_{ICME}
Full Set	2,617	1,702	915	474	256	228
R18 CMEs	720	514	206	138	84	83
New CMEs	1,897	1,188	709	336	172	145

columns show values related to number of predictions—the total number submitted, n_P , the number corresponding to impacts, $n_{P,I}$, and the number corresponding to non-impacts, $n_{P,NI}$. The next three columns represent numbers related to individual observed CMEs (i.e., there are typically multiple predictions per CME). These show the total number of coronal CMEs in the data set, n_{CME} , the number of coronal CMEs that actually have a corresponding observed impact, $n_{CME,I}$, and the number of in situ CMEs, n_{ICME} . Note that the final two columns are not identical, suggesting that sometimes more than one coronal CME is associated with each in situ event, which we discuss further below.

We note that R18 list their total number of predictions at 724, as compared to 720 listed here. Upon closer inspection we discovered that one event in the ATSB (arrival at 21 October 2017 03:00) was actually an impact at STEREO A and not Earth. We have removed the four predictions for this event from the revised study.

3.1. Non-Impacting Events

The full data set now has 2,617 predictions, over a 3.5 times increase from R18. Table 2 splits the predictions into cases with corresponding impacts ($n_{P,I}$) and non-impacting cases ($n_{P,NI}$) for which no corresponding in situ CME observations can be found. The ATSB is meant to collect predictions for all expected impacts at Earth, and we find that roughly 2/3 of the submitted results do correspond to impacts. It is not a collection of the expected behavior for all CMEs, hence we cannot perform a proper hit-miss analysis without collecting a complete list of all events and assuming that a non-submission to the scoreboard implies the prediction of a non-impacting event. It is interesting to note, however, that roughly a third of submitted predictions are false alarms.

These numbers all represent predictions, not individual CMEs. The n_{CME} and n_{ICME} columns give the corresponding number of coronal and in situ CMEs for these predictions. Interestingly, the number of in situ CMEs is much smaller (one half or less) than the number of coronal CMEs, even though only a third of the predictions were false alarms. We find that the 228 CMEs observed in situ correspond to 256 coronal CMEs. There are a number of cases where two or more coronal CMEs are associated with the same in situ event (e.g., the 1 April 2014 17:00:00 and 2 April 2014 13:55:00 eruptions are both linked to the 5 April 2014 09:40:00 arrival). The majority of these events have occurred within the new data set. CMEs certainly interact as they travel through interplanetary space (see, e.g., Lugaz et al., 2017), hence it is not unexpected to observe a combined structure at 1 au. Ideally, one would be able to separate the combined structure into multiple components with separate start times, but this is not always feasible and the ATSB generally does not include that level of detail.

Comparing $n_{CME,I}$ with n_{CME} , we find 218 coronal CMEs with no corresponding in situ counterpart. Looking closer at the data we find 78 of these correspond to events with a single submitted AT prediction (plus the “Average of all Methods” that is automatically generated, even for single event cases). About 75% of these non-impacting CMEs have four or fewer submitted predictions but we do see 21 events with 8–11 predictions. Overall, we conclude the majority of non-impacting cases were only simulated by a small fraction of the community, though there are a small number of interesting cases where a legitimately, by consensus, unexpected false alarm may have occurred.

3.2. Metrics

We calculate the error, Δt , as the difference between the predicted and the observed AT. In this section we present metrics for the overall properties of the errors. These are mostly common, straightforward measurements such as the mean, the MAE, and the standard deviation (σ). We do calculate some weighted metrics, which we weight using the confidence percentage (w_{CP}). Since the confidence percentage captures how likely it is that a CME will arrive, the weighted metrics are a rough filter on the AT error for more direct, rather than glancing/flank arrivals. The weighted mean (wMean) is calculated as

$$wMean = \frac{\sum_{i=1}^N w_{CP,i} \Delta t_i}{\sum_{i=1}^N w_{CP,i}} \quad (1)$$

where Δt_i is the error in prediction i and N is the total number of predictions in the set for which the wMean is being determined. The weighted MAE (wMAE) as

$$\text{wMAE} = \frac{\sum_{i=1}^N w_{\text{CP},i} |\Delta t_i|}{\sum_{i=1}^N w_{\text{CP},i}} \quad (2)$$

and the weighted σ ($w\sigma$) as

$$w\sigma = \sqrt{\frac{\sum_{i=1}^N w_i (\Delta t_i - \text{wMean})^2}{\sum_{i=1}^N w_i}} \quad (3)$$

For predictions where the confidence is not explicitly specified we assume 100% confidence. We also determine the minimum (Min), median (Med), and maximum (Max) values of the error and absolute error, as well as the first quartile (1Q, 25%) and third quartile (3Q, 75%) of the distribution. The values for the absolute error are identified with an “a” prefix on the metric name.

Table 3 shows all of the calculated metrics for the full data set, as well as split by the original R18 cases and the new data set. The results are also separated by individual models (using results from the full time span), which we will discuss in Section 6. The table lists the model name and $n_{\text{F,T}}$, which is the relevant number as the non-impacting cases do not factor into calculating the metrics. The table next shows the mean, wMean, MA, wMAE, σ , and $w\sigma$. The last two sections of the table show the Min, 1Q, Med, 3Q, and Max for the error and the aMin, a1Q, aMed, a3Q, and aMax for the absolute error.

We find a mean AT error of -2.5 hr, a MAE of 13.2 hr, and a σ of 17.4 hr for the full data set. If we consider the weighted values the mean slightly increases to -2.7 hr but the MAE and σ slightly decrease to 12.8 and 16.6 hr, respectively. We find that the weighting has minimal effect on the overall metric but we will explore its effect on individual model sets in Section 6.

Comparing the previous data set with the new one we see decreases in all of the metrics with the mean changing from -2.8 to -2.3 hr, the MAE from 14.0 to 12.8 hr, and the σ from 18.3 to 17.0 hr. These are suggestive of improvement, but it is unclear if these changes are statistically significant. When looking at the weighted values, we still see improvements but the changes are even smaller.

All three sets have cases with extreme errors of order three or more days (Min and Max), which is concerning considering the average transit time is of the same order. The middle 50% of predictions (between 1Q and 3Q) have errors within -12.3 and 8.0 hr for the full set, corresponding to a much more reasonable percentage error. The absolute errors range between zero, indicating at least one perfect time prediction, and the upper limit of several days that we see in the unsigned range. The middle 50% of the unsigned errors falls between 4.8 and 18.0 hr with a median value of 10.1 hr. There is a slight improvement between the original and new data sets, but again it is fairly marginal. In general, we are still predicting AT with a marginal bias of a few hours early and an average error of about 10 hr, but with a standard deviation of about 15 hr indicating significant scatter in the actual value of individual events.

4. Variations in Time

After reducing the full data set to a few summarizing metrics, we now look to stretch it out in different dimensions looking for any systematic variation in the error. We start with variations in time. Figure 2 shows a scatter plot of the signed errors (Δt) versus time. Each dot represents an individual prediction with the color indicating the model. The dashed line indicates zero AT error. This figure is directly analogous to Figure 1 of R18, though we have made no attempt to match model colors. R18 assigned colors from a traditional rainbow map using by the order in which models were added to the ATSB. Here we assign colors alphabetically from the “plasma” colormap in Python (at 50% transparency to better show overlapping points) with the darkest colors coming first and ending with the lightest. As such the yellow points correspond to the cluster of WEC models. We do not include a legend because it is impossible to distinguish the subtle differences in shade across 39 models. We note that the

Table 3
Metrics for the Arrival Time Errors

Model	$n_{p,1}$	Mean	wMean	MAE	wMAE	σ	$w\sigma$	Min	1Q	Med	3Q	Max	aMin	a1Q	aMed	a3Q	aMax
Full Set	1,702	-2.5	-2.7	13.2	12.8	17.4	16.6	-73.4	-12.3	-1.6	8.0	88.7	0.0	4.8	10.1	18.0	88.7
Riley Set	514	-2.8	-2.9	14.0	13.3	18.3	17.1	-66.9	-13.1	-2.4	7.5	69.5	0.0	5.2	10.8	19.7	69.5
New Set	1,188	-2.3	-2.7	12.8	12.5	17.0	16.4	-73.4	-11.8	-1.3	8.1	88.7	0.0	4.6	9.9	17.3	88.7
Anemomilos	19	-7.3	-7.3	22.0	22.0	26.2	26.2	-50.2	-22.6	-7.6	7.0	55.8	0.2	9.9	21.2	31.8	55.8
Avg. of all	256	-2.5	-2.1	11.9	11.0	15.8	14.5	-60.3	-11.1	-1.3	6.5	42.8	0.0	4.7	9.1	15.7	60.3
BHV	4	-1.3	-1.3	8.5	8.5	10.2	10.2	-17.5	-5.9	1.4	6.0	9.5	2.1	4.1	7.2	11.5	17.5
BGS	2	6.9	6.9	6.9	6.9	0.0	0.0	6.9	6.9	6.9	6.9	6.9	6.9	6.9	6.9	6.9	6.9
CAT-PUMA	7	3.3	3.3	9.6	9.6	10.1	10.1	-16.9	-0.4	5.7	10.2	14.5	4.6	5.5	8.5	13.2	16.9
CMEFM	21	-1.0	-1.0	13.7	13.7	16.6	16.6	-30.1	-10.8	-5.0	9.3	31.2	0.1	7.7	10.7	18.9	31.2
COMSEEP	4	19.7	19.2	19.7	19.2	8.7	8.9	8.5	12.7	20.6	27.7	29.2	8.5	12.7	20.6	27.7	29.2
Cone + HAF	26	2.6	2.8	10.4	10.3	12.0	11.9	-24.1	-6.3	4.0	10.8	24.0	0.3	5.2	9.7	13.3	24.1
DBM	15	5.1	4.8	11.1	11.0	12.1	12.1	-18.1	-1.5	5.3	11.7	26.0	2.0	5.2	8.9	16.8	26.0
DBM + ESWF	3	10.8	10.9	10.8	10.9	3.9	4.0	6.9	8.2	9.5	12.8	16.1	6.9	8.2	9.5	12.8	16.1
EAM	185	-6.3	-6.4	13.6	13.6	16.8	16.8	-63.4	-16.0	-4.0	6.0	36.4	0.0	4.9	11.1	18.1	63.4
ELEvo	5	4.4	4.4	7.5	7.5	8.3	8.3	-6.5	-1.2	4.8	6.8	18.0	1.2	4.8	6.5	6.8	18.0
ELEvoHI	2	46.3	5.4	46.3	5.4	42.4	11.2	3.9	25.1	46.3	67.5	88.7	3.9	25.1	46.3	67.5	88.7
ESA	3	-16.3	-16.3	18.7	18.7	14.6	14.6	-31.0	-26.3	-21.5	-9.0	3.5	3.5	12.5	21.5	26.3	31.0
EnsWEC (SWRC)	49	-7.8	-6.7	14.2	13.4	16.8	15.2	-64.8	-17.7	-5.7	4.2	24.4	0.6	5.4	11.8	18.6	64.8
EnsWEC (M2M)	95	-4.8	-4.3	10.3	9.9	12.7	12.3	-51.5	-14.4	-4.2	3.7	32.3	0.1	4.2	8.9	14.9	51.5
ESPM	2	-19.7	-19.7	19.7	19.7	8.2	8.2	-27.9	-23.8	-19.7	-15.6	-11.5	11.5	15.6	19.7	23.8	27.9
H3DMHD	1	-5.5	-5.5	5.5	5.5	0.0	0.0	-5.5	-5.5	-5.5	-5.5	-5.5	5.5	5.5	5.5	5.5	5.5
HAFv2w	1	1.8	1.8	1.8	1.8	0.0	0.0	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8
HUXt	3	14.5	15.3	14.5	15.3	7.2	7.7	5.6	10.2	14.8	19.0	23.2	5.6	10.2	14.8	19.0	23.2
ips.gov.au	2	-4.0	-4.0	4.0	4.0	1.5	1.5	-5.5	-4.7	-4.0	-3.2	-2.4	2.4	3.2	4.0	4.7	5.5
IZMIRAN	1	-6.1	-6.1	6.1	6.1	0.0	0.0	-6.1	-6.1	-6.1	-6.1	-6.1	6.1	6.1	6.1	6.1	6.1
NSSC SEPC	3	8.5	9.0	18.1	17.8	16.8	16.1	-14.4	0.0	14.5	19.9	25.4	14.4	14.5	14.5	19.9	25.4
Ooty IPS	2	-5.2	-5.2	12.7	12.7	12.7	12.7	-17.9	-11.6	-5.2	1.1	7.5	7.5	10.1	12.7	15.3	17.9
Other	2	-40.5	-9.0	40.5	9.0	33.0	9.8	-73.4	-56.9	-40.5	-24.0	-7.5	7.5	24.0	40.5	56.9	73.4
Rice-ENLIL Dst	1	-25.0	-25.0	25.0	25.0	0.0	0.0	-25.0	-25.0	-25.0	-25.0	-25.0	25.0	25.0	25.0	25.0	25.0
SAO Crowdsourc	2	-1.5	-1.2	1.5	1.2	1.0	1.0	-2.5	-2.0	-1.5	-0.9	-0.4	0.4	0.9	1.5	2.0	2.5
SARM	73	-7.7	-7.7	12.1	12.1	14.1	14.1	-54.0	-16.2	-6.0	0.9	21.1	0.2	3.6	9.4	17.7	54.0
SIDC	175	2.9	2.8	14.4	14.0	18.7	18.1	-56.9	-8.0	2.7	13.4	56.9	0.0	5.4	10.5	21.6	56.9
SPM	38	2.5	2.5	12.7	12.7	16.6	16.6	-32.7	-9.2	2.9	10.2	48.4	1.8	3.9	10.1	16.3	48.4
SPM2	40	-14.2	-12.0	19.4	16.2	22.7	19.8	-67.5	-22.3	-9.8	1.7	33.0	1.0	5.2	15.8	27.2	67.5
STOA	7	-21.5	-21.5	22.9	22.9	18.6	18.6	-55.3	-28.5	-24.7	-9.1	4.9	0.1	11.5	24.7	28.5	55.3
WEC	8	-7.1	-7.1	12.2	12.2	14.6	14.6	-28.7	-12.4	-6.9	0.4	19.7	0.3	5.0	7.4	21.5	28.7
WEC (BoM)	63	-0.6	-2.2	13.5	13.4	18.7	18.2	-54.8	-9.8	-0.5	10.0	49.8	0.2	4.1	10.1	17.5	54.8
WEC (SWRC)	99	-3.9	-3.7	13.9	13.8	17.5	17.6	-48.0	-15.0	-5.6	5.9	41.2	0.0	5.7	10.7	20.0	48.0
WEC (KSWC)	23	-7.2	-7.2	20.0	20.0	22.1	22.1	-47.3	-23.4	-9.5	12.8	35.9	2.0	11.7	19.8	28.2	47.3
WEC (Met Office)	162	0.5	0.0	13.7	12.9	18.5	17.4	-66.9	-9.9	1.8	10.2	69.5	0.1	4.6	10.0	18.0	69.5
WEC (M2M)	174	-2.5	-2.5	10.8	10.8	13.7	13.7	-48.9	-10.7	-2.0	6.2	32.5	0.4	3.5	9.1	15.3	48.9
WEC (SWPC)	124	0.1	0.1	12.8	12.9	16.3	16.3	-54.3	-9.7	1.0	10.0	35.0	0.1	5.1	10.0	19.1	54.3

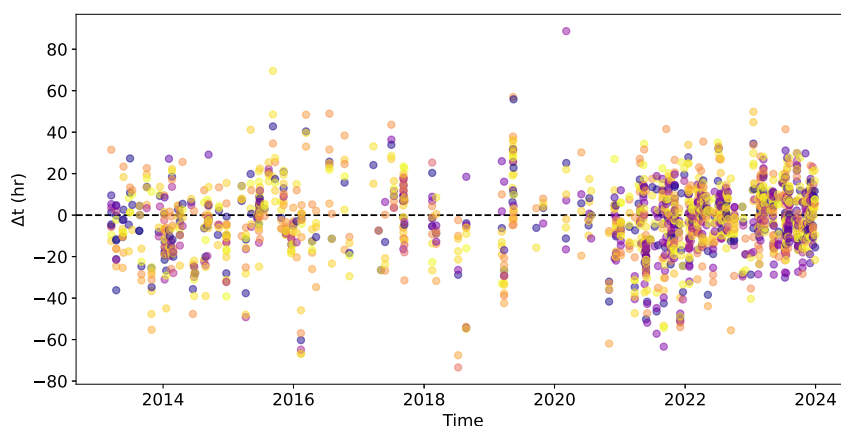


Figure 2. Scatter plot of the error in arrival time over time. Each dot is colored according to the model used for that prediction.

same color mapping is used in Figure 9 and there the individual points have numbers matching the model numbers used in Table 1.

The first half of the figure is identical to R18, beyond the difference in color. We see a decrease in the number of predictions between roughly 2018 and 2020, which corresponds to solar minimum, and fewer CMEs occurring. Around 2021, the number of predictions begins rapidly increasing, which is a combination of an increase in solar activity and an increase in the number of models having results submitted to the ATSB. The density of points has become so high that it becomes hard to infer much detail from a scatter plot due to the overlapping cases.

To address this overlap, we produce a heat map timeline of the AT errors, shown in Figure 3. We essentially transform Figure 2 onto a grid, discretizing both the time and error axes and summing the number of predictions in each grid cell. The color of a cell indicates that number of predictions, as indicated by the color bar. The top panel shows the signed error and the bottom shows the absolute (unsigned) error. The lines indicate rolling averages of either the mean (blue) or median (cyan) error over a 1 year window, centered at the middle of each time grid cell. Note the difference in the vertical range of each panel. We remove the negative portion for the unsigned errors and halve the positive range to better show the details at low errors. This leads to some accumulation in the highest grid cells for absolute errors of 30 hr or above.

Figure 3 highlights the obvious increase in the number of predictions since about 2021. We lose any information about the individual models but can better see the trends over time. The bias (signed error) remains fairly small over the full duration, typically remaining with ± 10 hr. In 2018, we see a brief dip in the mean signed error toward -20 hr, but this is also a low activity period with only a few observed events. Beyond about 2022 the bias remains very low, essentially consistent with the zero error line.

We find similar behavior to the signed error in the unsigned error. The initial value is moderate, around 10 hr as often quoted and shown in R18. The error begins to increase around 2016, reaching one peak near 2017 and a higher one around 2019. We note that these peaks are largely a result of the error increasing as the transit time increases for the slower CMEs during solar minimum, which we discuss further in Section 7. Beyond 2019, the error decreases and remains at roughly pre-2016 values until mid-2021, after which it further decreases by a few hours. As with the bias, this final decrease could signal a slight improvement in our predictive capabilities but it needs to be confirmed with additional measurements over the coming years.

5. Variations Between Individual CMEs

In this section we group the predictions by observed CME to search for any trends with respect to various CME properties. In the previous figures, we saw an increase in the number of predictions after 2021 so we first explore the balance between the growing number of CMEs and the number of predictions being submitted for each CME as the number of models has increased.

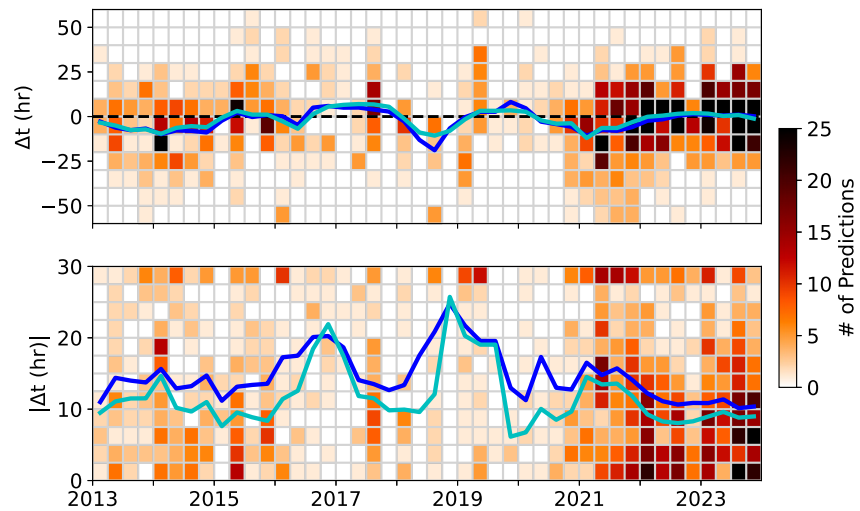


Figure 3. Heat map timeline of the errors in arrival time for all predictions. The top panel shows the signed error and the bottom panel shows the absolute (unsigned) error. The lines show a rolling mean (blue) or median (cyan) over ± 6 months.

5.1. Number of Predictions Per CME

Figure 4 shows a scatter plot with each dot representing a individual CME and the y-axis position showing the number of predictions for that event and the x-axis showing time. Each dot is colored by the MAE (unsigned error) for the predictions for that event.

We see that early on there were certainly fewer predictions per event with only two events having more than 10 predictions before 2017. As solar minimum approaches the number of CMEs decreases, but the number of predictions per event trends upward. This is partially from new models being developed, but we also suggest that the more isolated solar minimum CMEs may be more appealing to simulate with simpler models that may struggle during high levels of activity increasing both the number and complexity of events. Figure 1 does show a handful of models that begin submitting predictions around 2018 then only have sporadic (or even no) predictions as solar activity begins increasing (e.g., BGS, CAT-PUMA, DBM + ESWF, ELEvo, ips.gov.au, NSSC SEPC, and Ooty).

Around 2021 activity begins to pick up and we find a wide range in the number of predictions for individual events. We see a maximum of 18 predictions for the 2 November 2021 02:48:00 UT CME eruption. To better quantify the results of Figure 4 we determine the mean number of predictions per CME per year. From 2013 to 2016 the mean ranges between 5.2 and 6.4 predictions per CME whereas from 2021 to 2023 it ranges between 6.6 and 7.0 predictions per CME. This suggests a systematic increase in the number of predictions per event when comparing only the times of higher solar activity.

We note that the consistent six or so predictions predominantly come from the operational centers that routinely forecast nearly every CME. The events with a significantly higher number of predictions must somehow pique the interest of the general scientific community. We remark that individuals subscribed to the scoreboard mailing list are notified by email every time a new CME is added to the ATSB, inviting them to submit their prediction. One may then wonder why these particular CMEs appeal to the broader audience. The color of the points in Figure 4 shows no evidence of better predictions (lower errors) for the popular events. There are not any significant trends in general with respect to the number of predictions and the AT error.

5.2. Variation With Transit Time

We next investigate if there are any trends with respect to the intensity of an event—whether it is an extremely fast, rare event or a relatively common slow event. Given the information within the ATSB, our only measure of this is the transit time (hereafter TT) of the CME (calculated between the coronal CME time and the shock AT). Drag-like interactions will affect the exact TT beyond a simple linear relation between speed and time, but faster CMEs should still arrive quicker than slow ones. Figure 5 shows the mean error (left panel) and MAE (right) for

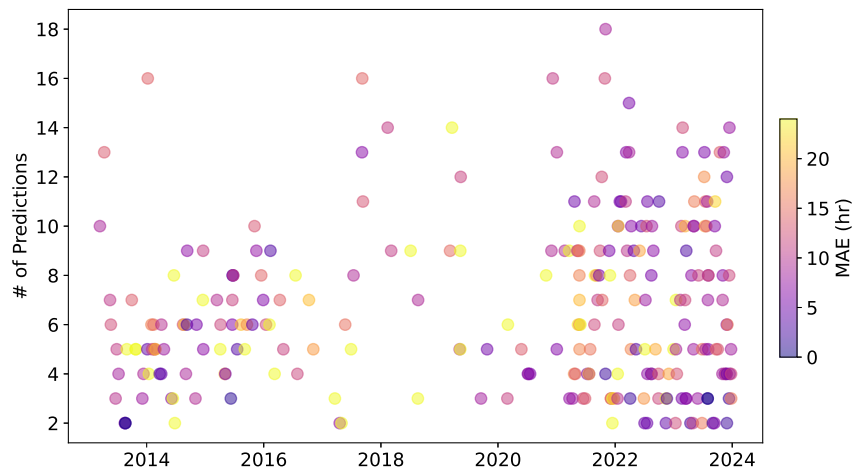


Figure 4. Scatter timeline of the number of predictions for each individual observed Coronal mass ejection. Each point is colored by the unsigned arrival time error.

each CME versus its TT. Each dot is an individual CME and it is colored by the number of predictions for that event.

We see that the events with the highest number of predictions tend to have the shortest TT, though there is quite a bit of variation. This supports the idea that the popular events tend to be the more “exciting,” extreme ones. More importantly, we see a strong negative correlation between the mean error and the TT. This means that predictions tend to be too late for short TT CMEs and too early for long TT CMEs. The correlation is strong with a Pearson coefficient of -0.569 (p -value of 10^{-23}). This suggests that the models are altering the transit speed too much toward the background solar wind speed, causing extreme events to propagate slower than observed and slow events to move faster than observed.

If we focus on the higher number of prediction cases or look at the points separated by model (not shown) we see that the popular events show less of a trend than seen in the left panel of Figure 5. Up to a TT of about 100 hr the most popular events tend to fall between ± 25 hr error with a slight hint of a negative correlation. Above 100 hr TT, the correlation between TT and error appears with longer TT having more negative errors. When separated by model we see that the WEC variants all tend to fall in the ± 25 hr band, suggesting the systematic variation is more strongly driven by other types of models.

We see a correlation between the MAE and transit time in the right panel of Figure 5, but it is weaker with a Pearson coefficient at 0.406 (p -value of 10^{-11}). Here we see that the larger TT have higher MAE. If we instead consider a mean absolute percentage error by normalizing the error by the TT (not shown) then the correlation

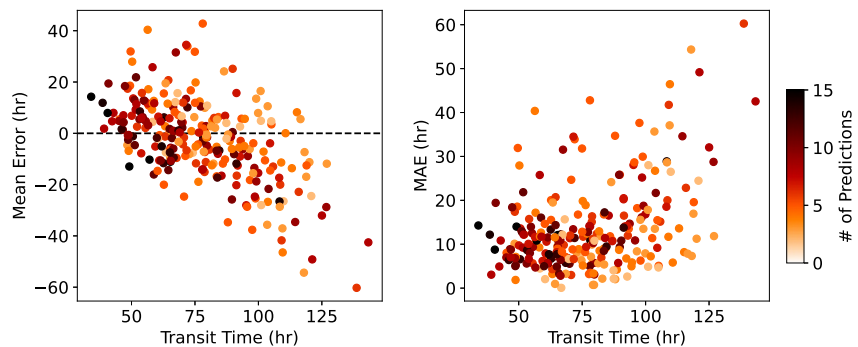


Figure 5. Scatter of arrival time error versus transit time for each coronal mass ejection, colored by the number of predictions for that event. The left panel shows the mean error and the right the mean absolute error over all predictions. In the left panel the dashed horizontal line indicates zero error.

drops to -0.066 and erases any further systematic trends with the TT. The majority of points fall below a 30% absolute error. Four events have a error above 50% but these all have fewer than 5 predictions each.

5.3. Variation With CME Properties

The ATSB does not include CME properties, but the CCMC also hosts the Space Weather Database Of Notifications, Knowledge, Information (DONKI, <https://kauai.ccmc.gsfc.nasa.gov/DONKI/>), populated in real time by the M2M Space Weather Analysis Office, that includes a reconstructed coronal latitude, longitude, angular width (AW), and speed for each observed CME. There is not an exact one-to-one correspondence between the identification (ID) tags used within DONKI and within the ATSB, but the small number of differences can easily be spotted and coordinated between catalogs. These mismatches typically only occur when a user submits an ATSB prediction before the coronal CME is officially added to DONKI, and the difference is only a few minutes or hours change in the time in the ID tag.

We use the DONKI catalog to find the latitude, longitude, AW, and speed of the coronal CME corresponding to each impacting event within the ATSB. Figure 6 shows correlations between the errors (mean on left, MAE on right) and the reconstructed CME properties. From top to bottom Figure 6 show the absolute latitude, absolute longitude, AW, and velocity (v). Each dot represents an individual CME and is colored by the corresponding number of predictions for that event. We show the absolute latitude and longitude because we expect any errors to depend on absolute distance and not the particular direction.

Unlike for the transit time, we see no evidence of any significant correlations. The strongest correlation is between the MAE and the velocity at a Pearson r of -0.166 , but this is a p -value of only 0.008 and not significant. We might have expected to find better predictions for more Earth-directed events (low absolute latitude and longitude) but there is no evidence for that. The AW tends to scale with CME intensity, so we expect the biggest events to be the fastest and therefore have the shortest TT, and we would expect the same trend as seen in Figure 5 but we find nothing. We do not even see a significant trend when comparing directly to the reconstructed CME velocity.

This is quite surprising given the strength of the correlation in Figure 5. We also note that W18 found a strong correlation between speed and error for the WEC (GSFC SWRC) results at both Earth and the STEREO satellites (Figure 3 of W18) but it actually shows the opposite of our TT trend. W18 find a trend of predicting early arrival for fast CMEs, whereas we see late predictions for short TT and no trend with speed. We suspect any trends with CME parameters may not appear in our results because we are associating the DONKI parameters to the average errors from all predictions. The DONKI values should be used for the NASA GSFC predictions and should be similar to what was used by others, but each forecaster likely has their own coronal CME reconstruction that then use to drive their model. Coronal reconstructions are notoriously uncertain (e.g., Kay & Palmerio, 2024; Verbeke et al., 2023) so we expect there could be significant variation between the inputs used by different models. We have no means of comparing with the specific values used for each prediction, but suspect we may find more significant correlations if we could. We also have no explanation for why our TT trend differs from the velocity trend found by W18. We acknowledge that we are looking at the collection of results from many different models whereas W18 was focused only on one model from one forecasting team. We still find no correlation if we just compare the WEC (GSFC SWRC) errors and the CME speed, but the data sets are still not the same since we restrict to only near-Earth events and include several additional years as compared to W18. It is also possible that after reviewing the bias error trends in W18 for fast CMEs, GSFC adjusted their CME measurement procedures for the speed and width.

6. Variations by Model

Our final manner of separating the results is by model, with model still referring to the name of a set and may include both the type of simulation that was used to generate the results and the team that used it. Table 3 lists the metrics for each model, listed in alphabetical order. The table contains an overwhelming number of measures but provides a thorough, unbiased characterization of each model so we include it for reference.

The mean values vary between -40.5 and 46.3 hr, the MAE between 1.5 and 46.3 hr, and the σ between 1.0 and 42.4 hr (ignoring zero σ values for models with only one prediction). As with the full data set, we see little difference between the unweighted and weighted metrics for most models but there are notable exceptions. One example is the ELEvoHI set. There are only four submitted predictions, only two of which correspond to impacts.

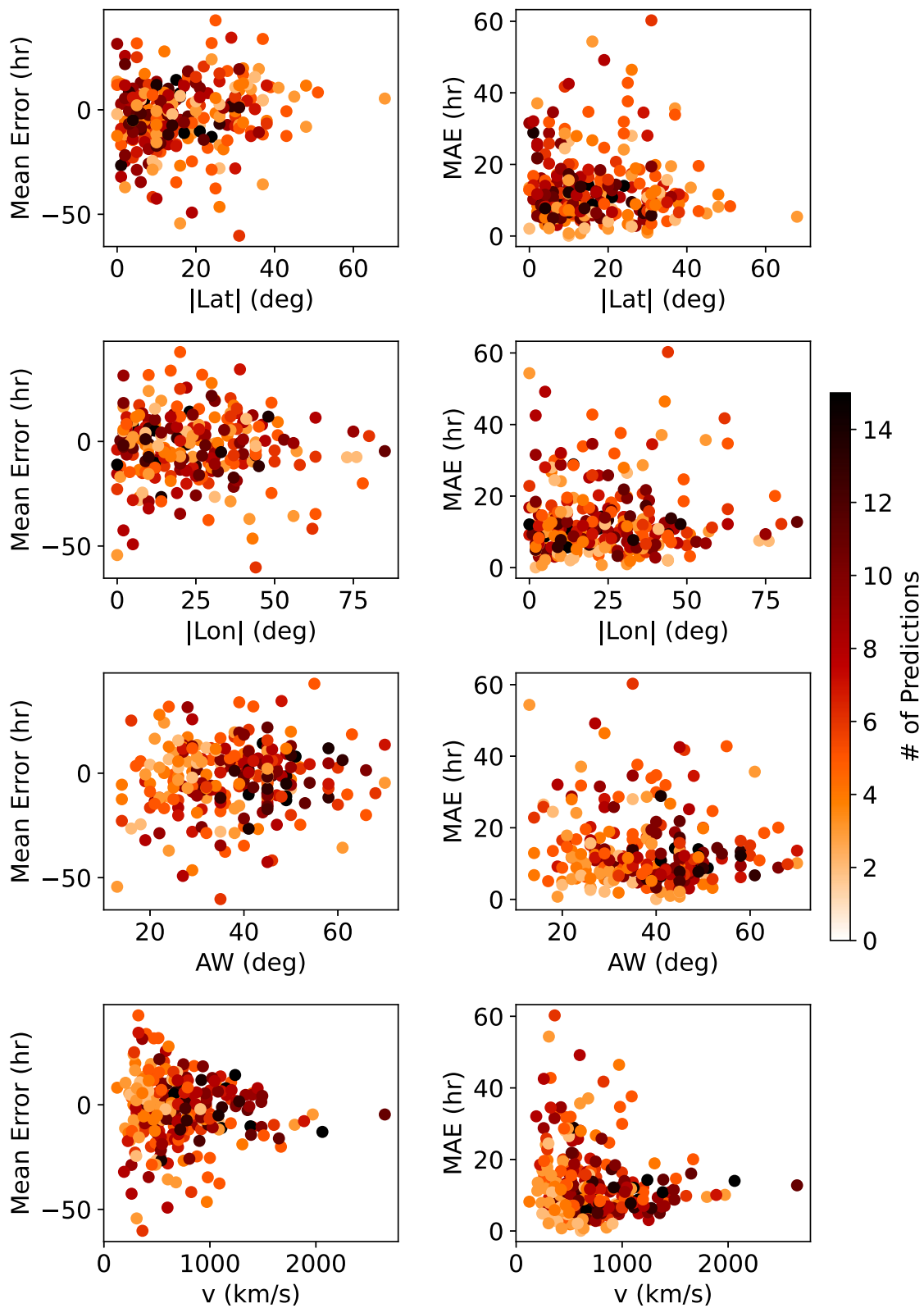


Figure 6. Scatter plots of the error versus different coronal mass ejection (CME) properties. The left column shows the mean error and the right column the mean absolute error. From top to bottom, each row shows the absolute CME latitude, the absolute CME longitude, the angular width, and the coronal CME velocity. The points are colored by the number of predictions for that CME.

One of these events (2 March 2020 20:00:00 coronal identification, 6 March 2020 14:00 in situ arrival) is a flank encounter with difficult-to-interpret in situ signatures. The ELEvoHI prediction ended up with an error of 88.72 hr, but only had a confidence percentage of 1.8%. As such, including the weights greatly improves the ELEvoHI metrics, but this is by far the most extreme example of the effects of weighting. The weighted means range between -25.0 and 19.2 hr, the weighted MAE between 1.2 and 25.0 hr, and the weighted σ between 1.0 and 26.2 hr. Most individual models that include weights and have more than a few predictions have their metrics only change by about an hour or less but we encourage the benefit of adding confidence levels, particularly for highly uncertain events.

6.1. Largest Model Sets

In general, we find that the most extreme errors tend to come from the models with smaller sample sizes (i.e., a smaller number of submitted predictions). To better characterize the more common predictions we focus on the most-frequently submitted models (which we will refer to as “most popular” for simplicity). R18 performed a similar analysis, highlighting the results from the six most popular models. These were, in order of most predictions, the Average of All Methods, WEC (GSFC SWRC), SIDC, WEC (NOAA/SWPC), WEC (Met Office), and Ensemble WEC (GSFC SWRC). Here we include the 11 largest sets, a strange cutoff number but chosen to make sure we include all the models that were included in the deeper analysis in R18. In addition to the previous six from R18, we include EAM, WEC (NASA Space Weather Analysis Office), Ensemble WEC (NASA M2M), SARM, and WEC (BoM). This represents all models with 90 or more predictions and 63 or more impacting cases.

Figure 7 shows histograms of the unsigned error for each model, organized from left to right then top to bottom by number of predictions. Each histogram is colored like the corresponding model in Figure 2. The vertical black line shows zero error and the red dashed line shows the mean of that distribution. The dotted red lines indicate $\pm 1\sigma$ and the red curve is the normal distribution corresponding to that mean and σ . We also include the normal distribution from R18 as a blue line for the six models that were highlighted in that work. We have scaled the peak to match that of the new data to facilitate comparison. All x -axes share the same range but note the difference in scale on the y -axes.

We can visualize the bias within each model by comparing the vertical dashed red and black lines. When they overlap there is essentially no bias. If the red line falls to the left of the black then there is a tendency for early predictions and to the right corresponds to late predictions. Most models have a small bias and tend more toward early predictions, which matches the trends of the full data set. We see a few models with a late prediction bias, but the magnitudes of these positive biases are quite small. For these 11 most popular models, the biases range between -7.7 and 2.9 hr with the smallest absolute bias of 0.1 hr for WEC (NOAA/SWPC).

There is a small difference between the width of the distribution across different models. Table 3 shows the differences in σ but for the most part it is not particularly distinguishable in Figure 7 comparing either the normal distributions or the location of the dotted lines. For these models we find σ between 12.7 and 18.7 hr and the MAE between 10.3 and 14.4 hr. All of these metrics are comparable to what we see for the full data set, but this is not surprising since these 11 models make up 85.5% of the full results.

For the six cases highlighted in R18, we see either minimal change or an improvement in the AT capabilities based upon comparison of the previous and revised distributions. Most noticeably, the bias improves for the Avg. of All set and the width significantly decreases for WEC (Met Office) and to lesser extents for all the other models. We cannot compare the NASA M2M Space Weather Analysis Office results to previous results as it did not exist at the time of R18, but we find the NASA M2M Space Weather Analysis Office results (both single case and ensemble) are improvements relative to its predecessor GSFC SWRC.

To further illustrate the abilities of these 11 models relative to the whole data set we show the 6-month rolling mean of each model over time in Figure 8. The figure has a similar format as Figure 3 but we only show the rolling mean, not the heat map or the rolling median. The blue line shows the value for the full set, the exact same as in Figure 3. The other lines represent other models, using the same colors for the points as in Figure 7 and connected with a thin gray line for visual clarification. The left column shows the mean error and MAE in hours and the right column uses the transit time of each CME to normalize these metrics before taking the rolling average.

For the most part, these models fall fairly symmetrically about the full catalog values. We see a few time periods with extreme outliers in an individual model, most noticeable for an orange case in 2018–2019. This corresponds

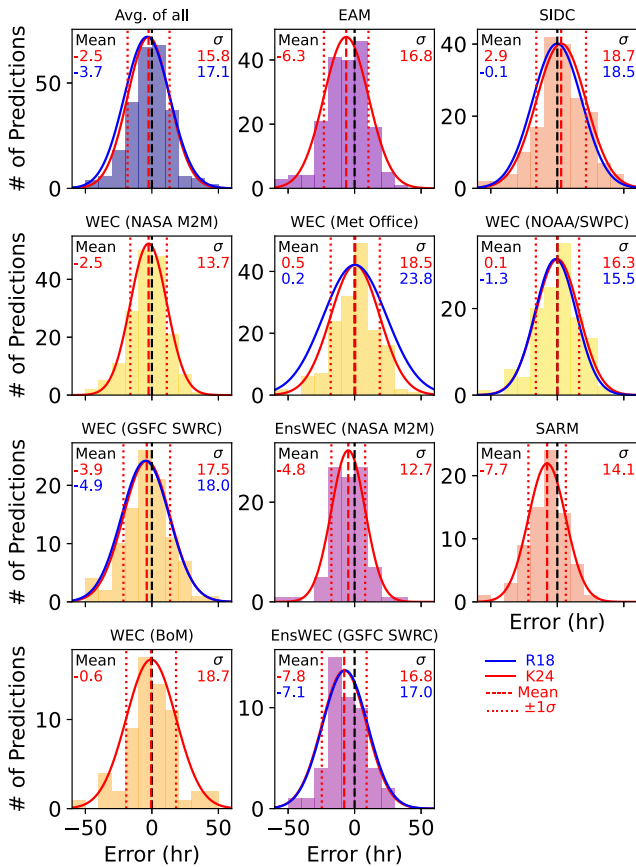


Figure 7. Histogram of the errors for the 11 models with the highest number of predictions. The black dashed line is at an error of zero. The red dashed lines are at the mean value of each distribution and the red dotted lines at plus or minus one standard deviation from the mean. The title of each panel shows the model name and the red lines show Gaussian fits to each set. For models included in R18 the blue line shows the Gaussian fit from the previous work. The models are colored in the same manner as Figure 2.

to the WEC (BoM) model and solar minimum times. Over nearly a 1.5 yr span they have only a single prediction with corresponding impact, which is for the 19 August 2018 07:54:00 UT CME. This prediction has an AT error of -55 hr and is the only event determining the rolling mean for two of the discrete time steps shown in Figure 8.

We also see a brief time near 2020 when many of the models consistently had a better unsigned error than that of the full set. This is still in solar minimum when there were few CMEs, meaning both fewer events to predict and that those events should be easier in terms of fewer interactions with other CMEs or high speed streams. It appears that these popular models are capable of accurately reproducing the CME propagation during this time, which implies that the less-popular models must be introducing larger errors to drive up the full set value. We suggest this may be a sign of more models being added into the ATSB at this time.

When the errors are expressed in hours we find an increase in the error around solar minimum. However, we know that there tend to be more slow CMEs with long transit times during solar minimum and that our data has a strong correlation between error and transit time. If we consider the normalized errors in the right column of Figure 8 we see that there are still peaks in the bias, alternating between late and early predictions, but the magnitudes are comparable to what we see at earlier times. The normalized bias at the end of this time trends toward slightly late predictions by about 5%, which would represent improvement from the earliest predictions if it continues at this rate.

The unsigned error in hours also shows the increase in error near solar minimum for both the full data set and the individual models. When normalized, the peaks in the MAE at solar minimum are significantly reduced. We see errors near 20% at the start of the study, followed by an increase to about 25% between 2017 and 2020, followed by a decrease to about 15% for the remaining time. Both the MAE and normalized MAE show a wide range in the individual model lines about the full set curve at early times. During the late period the model lines are much more tightly grouped, showing far less variation about the full set behavior. Combining this with the improvement in the distributions from Figure 7, we infer while the

improvement in the models, both in bias and uncertainty, is small and of questionable statistical significance, we certainly see an improvement in the consistency of the models with the majority of the most popular models performing fairly similar by the end of 2023.

6.2. Comparison of All Models

We extend our analysis of the relative ability of the different models to include all of them submitted to the ATSB, not just the most popular ones. Figure 9 shows the mean (left) and MAE (right) versus the average lead time for each model. We define the lead time as the temporal difference between when the prediction is submitted and when the CME arrives. The individual dots represent the combined values over all impacting CMEs predicted by that model. The symbol size is based on the number of predictions with larger symbol size indicating a larger sample size. The symbols are again colored by the model name but the small number matches the first column of Table 1 so the individual points can be identified. Note that this figure does not include the Average of All model because it is not given a lead time in the ATSB.

We see that most models have a small bias and fall close to the dashed line indicating a mean error of zero in the left panel of Figure 9. The most popular models all fall in a cluster near the zero error and between about 40 and 60 hr lead time. The smaller models sets show a bit more bias than the larger sets, and many also have shorter lead times. We find no significant trend in the bias with respect to the lead time. The largest lead time does have the largest (most negative) bias, but this for a prediction from IZMIRAN, which only submitted two total predictions over the full time span.

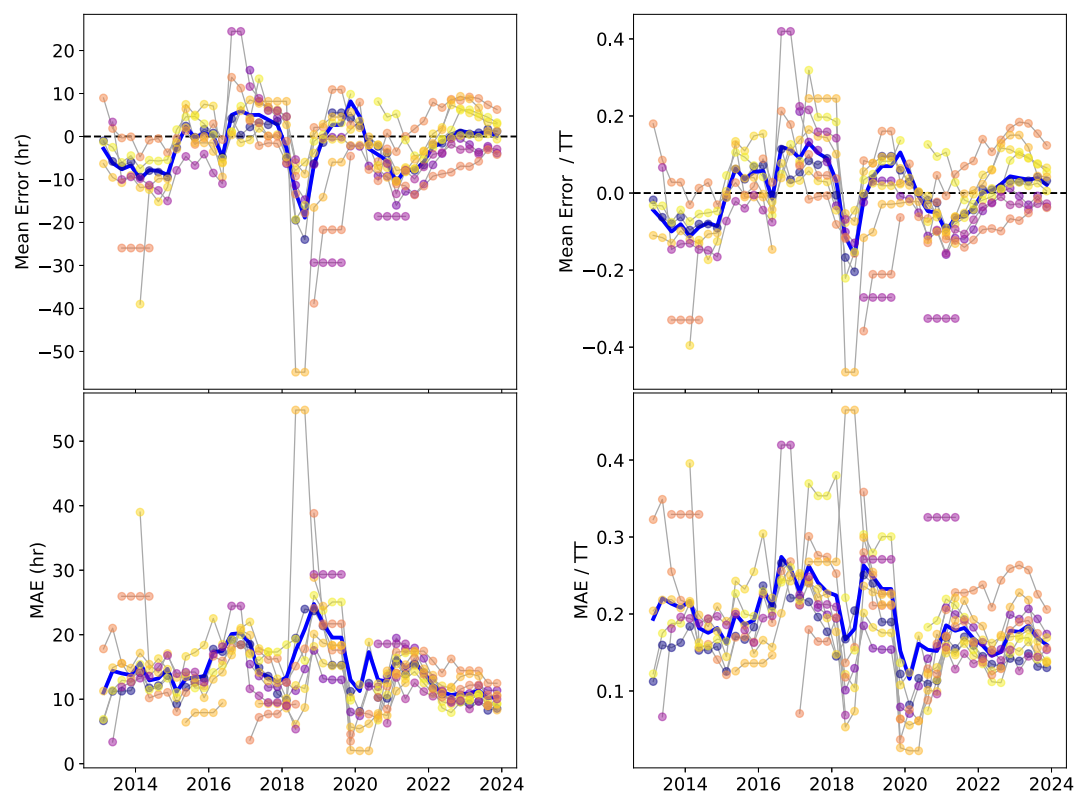


Figure 8. Rolling mean of the error (top) and absolute error (bottom) for the 11 most popular models. The left panels show values in hours and the right panels are normalized by the individual coronal mass ejection transit times. The blue line represents the rolling mean over all models, the same as in Figure 3. The points from the same model are connected by a thin gray line and the dots colored by the model name as in Figure 2.

Looking at the MAE, we again see the most popular models generally clustered together but more variation in the smaller sets. There is some correlation between a smaller error and shorter lead times, which would follow logically if we know the CME has not yet arrived and we gain more information or better constraints on CME properties as it continues propagating and further analysis can be done. We caution, however, that the majority of the trend is driven by the smaller sets and this correlation disappears when only considering the popular models.

We wish to define a single metric quantifying the success to facilitate comparison between models. The bias represents any systematic errors toward early or late predictions. The MAE and σ both give a measure of the spread in the errors, but with the MAE representing the spread about zero error and σ representing the spread about the bias point. In general, we have shown the MAE in our figures as opposed to the σ because it is a better measure of how far a prediction is from the true arrival. When combining multiple metrics into a single value we opt to use σ instead of the MAE because it is a more pure measurement of the spread in the results and not affected by the bias.

We define our overall metric as the product of the absolute value of the bias times the spread. We calculate this as the absolute value of the mean multiplied by σ , with two caveats. First we set a lower limit of 1 hr on the absolute mean, which prevents an overemphasis on the bias, particularly for the models with a low number of events. Second, σ cannot be determined for the single event cases so we replace it with the MAE for those models. Figure 10 shows the mean (left), σ (middle, with MAE for the single event sets), and our overall metric (right) versus the total number of predictions for that model. The points are colored according to lead time (setting the average at the lower limit) and labeled with the model number from Table 1. In this figure the y-location of a point is a pure error metric but the x-location (sample size) and color (lead time) are both important factors in determining model quality, but with a far more subjective interpretation of their relative importance. We note that we do use the weighted mean and σ , which makes little difference for most models but is kinder to the few small number models that do include their confidence and have large errors on low confidence predictions.

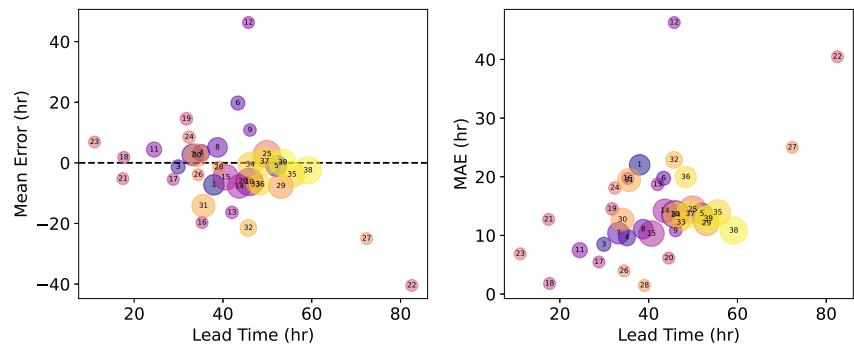


Figure 9. Error versus lead time. Each dot represents the results of a individual model, with the mean arrival time error shown on the left and the mean absolute error shown on the right. The symbols are colored by the model name and the number matches the first column in Table 1. The symbol size is determined by the number of predictions submitted for that model.

Again the most popular models tend to all cluster together in all three panels. The smaller sample sets show much more variety in their metrics. We do see the smaller sets tending to have smaller σ , which we propose could result from having less egregious outliers in the AT error when only considering a small, hand-picked set of more idealized events.

The combined metric shows that SAO Crowdsourc (label 27) has the best score at 1.0, but there are only two impacting events in this set (and one non-impacting). Of the larger sets (100+ cases), WEC NOAA/SWPC (39) and WEC Met Office (37) perform best with scores of 16.3 and 17.4, respectively. These are followed closely by SIDC (29) and WEC NASA M2M (38) at 32.8 and 34.2, which overlap in Figure 10. The 11 models highlighted in Section 6.1 all have combined metrics of about 100 or below. The worst score is for Rice-ENLIL Dst (26), but this is a single event set with an error of -25 hr leading to a combined metric of 625.

Overall, we find that the predictions from the well-established models that frequently submit to the ATSB are all fairly comparable in terms of quality. Some of the newer, or less frequently submitted models may be capable of more accurate results, as suggested by their metrics so far, but this remains to be confirmed with more events.

7. Discussion

The ATSB combines results from a variety of models, and not all of these models simulate the same properties or set the AT using the same definition. Some of the models predict the shock AT whereas others provide the CME ejecta/flux rope AT. All of these values are compared to the observed shock AT determined by the M2M Space Weather Analysis Office, with assistance from ICME experts for particularly complex events. The use of shock arrival times could produce a bias toward late arrivals for the flux rope models. Comparing those models identified as CME AT models, we find a mix of both early and late biases. Disentangling the contribution of any

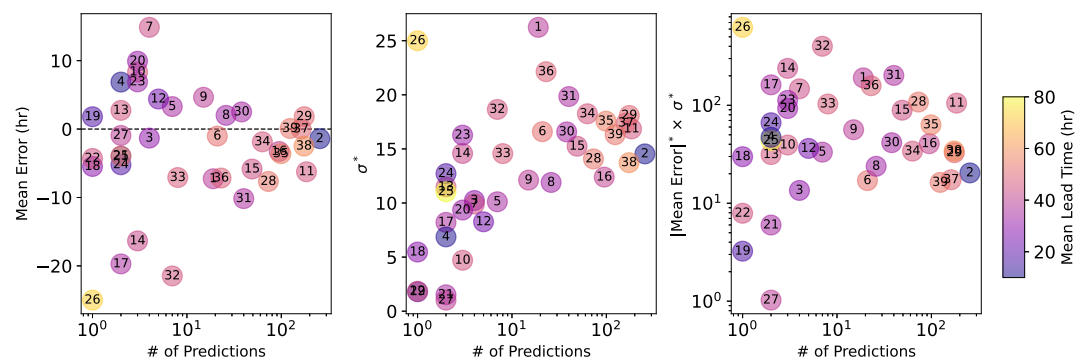


Figure 10. Scatter plots of the metrics for each model versus the number of predictions made with that model. The panels show the mean, mean absolute error (MAE), and our combined metric (absolute value of mean times the MAE) from left to right. Each point is colored by the average lead time for predictions from that model and the points are labeled by the model number.

systematic errors from the model itself versus the association of shock to ejecta ATs requires further analysis beyond the scope of this work.

A portion of the results provide error bars on the predicted time for individual events, which we can compare with the observed metrics. Ten different models have error bars included for 20 or more events. For these models, the observed error falls within the given range only 35%–55% of the time. For nearly all models the error bars are smaller than the corresponding MAE of that model. Both factors highlight the need to improve our quantification of the uncertainty in the model results.

We tend to see the worst errors in solar minimum. This is interesting as we would expect these cases to be the easiest to simulate accurately as they are less complex in terms of interactions with other CMEs or other structures. Comparing the errors for isolated versus interacting events would be a worthwhile endeavor, but it is beyond the scope of this work. The solar minimum events also tend to be slower and smaller events, so they should have less inertia than the extreme events and be influenced more by background forces. However, the extreme events have larger velocity differentials with respect to the background so we would expect stronger forces in those cases. Additionally, the weaker solar minimum events tend to be fainter and more difficult to measure. This makes the measurement very uncertain and can introduce errors to the model(s).

We do see that most of the trend disappears if one considers the percentage error, relative to the transit time, which is not surprising given the correlation we found between the error and the transit time. The trend does not fully disappear when accounting for the transit time, so there must be some additional source of error. The larger errors may not result from individual models that routinely predict every CME actually doing worse for those CMEs, which is supported by the number of these models falling below the full set curve in Figure 8. Rather it may be the contribution from more models during these times, particularly newly-introduced ones, combined with a smaller number of events overall.

We do see a strong correlation between the lead time of individual predictions (not shown but the same as Figure 9 without grouping by model) and the error, but this is likely just repeating the correlation we found with the transit time. The lead time is simply the transit time minus a “delay” time between when the CME is observed in the corona and the prediction is actually submitted. If we consider the errors versus this delay time we find no significant correlations, supporting the lead time trend being driven by the transit time trend.

Another factor to consider is also the location of the satellites observing the coronal CMEs over this time range. In 2013 both STEREO satellites were approaching the far side of the Sun, then contact was lost with STEREO-B in October 2014. In August 2019, STEREO-A had reached quadrature with Earth and an optimal position for reconstructing earthbound CMEs (when combined with near-Earth observations). By November 2022, STEREO-A was within 10° longitude of Earth, essentially reducing observations to a single viewpoint, which is known to increase the uncertainty in the reconstructed parameters (Verbeke et al., 2023). Oddly, the general variation in the errors tends to vary inversely with the quality of the available stereoscopic observations. Clearly more factors than just the satellite viewpoints affect the AT errors. We do not have enough data from different viewpoints at different times of the solar cycle to fully disentangle the various effects. It is encouraging however, that we are currently in a period with the best-to-date error metrics despite have sub-optimal viewing conditions.

Had we found systematic errors, then we could have used this updated analysis of the ATSB to suggest specific steps that could help improve AT predictions. Unfortunately, we find very little systematic behavior, with only the exception of the correlation with TT. The fact that our current errors tend to be of the order 10 hr suggests that we at least understand the general nature of interplanetary CME propagation and do not need to entirely dismantle the current approach in favor of a new one. Improving our predictions then comes down to two factors—improving the physics in the models and/or improving the inputs we feed into the models.

In terms of the physics, we see a wide range of model types with the ATSB. The best-performing models do tend to use WEC, a 3D MHD simulation, and the most sophisticated and detailed physics-driven approach. However, EAM, an empirical model, obtains nearly as accurate predictions using substantially less physics, and requiring far less computational resources. On top of using different models, every prediction is developed using a specific group's own set of input parameters. Fully separating the effects of different types of models would require a study using uniform inputs. Without this we cannot comment further on the importance of the physics implemented in each model. CCMC does have plans to collect more model run metadata with each submission, via a json submission file that can be sent in automatically, which may allow for deeper analysis in the future.

The model inputs include both the reconstructed CME properties and the background solar wind through which the CME propagates. We know that CME reconstructions can be affected by large uncertainties (e.g., Kay & Palmerio, 2024; Verbeke et al., 2023) and that these can translate into AT errors of the order of 10 hr (e.g., Kay et al., 2020). This alone is sufficient to mask the effects of any background SW effects or inadequacies in the models. This does not mean that these factors do not also significantly contribute, but we cannot better quantify them without systematically reducing the CME reconstruction uncertainties. The most likely route toward improvement is from consistent stereoscopic observations, which may be achieved in the early-to-mid 2030s when the Vigil (Palomba & Luntama, 2022) mission reaches the Sun-Earth L5 point. Until then, we will hopefully see improvements as STEREO-A begins moving away from Earth and into more ideal viewing angles. STEREO-A should reach the L4 point around 2026, near solar maximum times as opposed to the rising phase when it was there in 2011 on its first lap around the Sun. This should provide a greater number of events to thoroughly test whether our techniques have improved when an optimal viewing angle is available.

Perhaps the easiest target for improvement is in the background SW. The majority of simple drag-based models use a very basic representation of the ambient wind, such as a radial profile that varies only with distance. The real SW has structure due to the interaction of the slow and fast wind (see, e.g., Richardson, 2018). Many studies have shown that CME propagation varies in different background SW regimes (e.g., Odstrčil & Pizzo, 1999a; Odstrčil & Pizzo, 1999b). Adding any sort of structure may be a quick way of improving the simpler physics-driven AT models. Fluid and MHD models do typically include the expected SW structure and efforts are underway to validate the accuracy of their results at 1 au (Reiss et al., 2023). To properly simulate the interaction with a CME throughout interplanetary space we emphasize that it is also important to validate these models closer to the Sun, using observations made for example, by Parker Solar Probe and Solar Orbiter. We note that the Vigil mission should improve the accuracy of the ambient SW model as it will provide magnetograms, which typically drive SW simulations, from the L5 perspective before the corresponding regions rotate onto Earth's view.

While we cannot directly point to specific sources of errors or quantify how much improvement we could make, we do have hope for future predictions. In addition to continued fundamental research into the structure and morphology of CMEs and their interaction with the solar wind, upcoming missions such as Vigil will certainly improve our CME reconstructions and modeling of the background SW. We may even see improvements as STEREO-A continues to orbit the Sun. We also note the importance of upcoming missions that are less focused on operational space weather and more on novel techniques for observing the Sun, such as the Polarimeter to Unify the Corona and Heliosphere (PUNCH; Deforest et al., 2022) mission. Novel approaches to how we view CMEs and the solar wind and how we process those observations could certainly lead to unexpected scientific breakthroughs that also result in improvements to our space weather forecasting capabilities.

8. Conclusions

We have used the predictions submitted to the AT Scoreboard to revise current estimates of AT errors. This work builds upon that of Riley et al. (2018), which used the same predictions up until May 2018. Here we extend the analysis to include six additional years (until the end of 2023), representing a sample size 3.5 times the size of the original work. In general, we find nearly the same metrics as before, with a mean error (or bias) of -2.5 hr, a MAE of 13.2 hr, and a standard deviation of 17.4 hr. If we split the data set between the previous predictions and the new ones we do see a slight improvement in each metric, but the difference is small.

We try to disentangle any correlation of the errors with CME properties but mostly cannot find any significant trends. The only significant correlation is with the CME transit time, with longer transit times corresponding to earlier predictions and shorter transit times to later predictions. Comparing directly to the CME velocity shows no such trend, but we acknowledge that the CME velocities we use (from the DONKI catalog) may not be the same values used for each of the individual predictions.

We look for any evidence of any actual improvement over the 11 year time span covered in this work. We do see trends in the bias and uncertainty over the solar cycle, with the largest errors occurring toward solar minimum (roughly 2017–2020). This is a somewhat counter-intuitive result as we would expect CMEs to be easier to simulate during this time as they tend to be more isolated and less complex events, though simultaneously they are fainter and harder to accurately reconstruct. They are typically slower events with longer transit times, which we know correlates with the error. If we normalize the errors by the transit time it removes most, but not all, of the increase in error at solar minimum. We suggest any remaining effects are due to either difficulties in

reconstructing the faint CMEs or potentially the result of more models contributing predictions during this time period and the effects of a smaller number of events in general. We do see more noticeable improvements of a few hours in both the bias and uncertainty if we compare the earliest predictions (2013–2015) with the most recent ones (2022–2023), which is roughly comparing similar levels of solar activity. The percentage bias changes from about -10% to 5% and the percentage MAE from about 20% to 15% .

When we look at the results of individual models we see fairly similar metrics between the models with large data sets. These are mostly large forecasting centers than tend to run some version of the WSA-ENLIL + Cone (WEC) model, but it does include a few other models. These “popular” models have biases between -8 and 3 hr, MAE between 10 and 14 hr, and standard deviations between 13 and 19 hr. In comparison, the full catalog has biases between -41 and 46 hr, MAE between 2 and 46 hr, and standard deviations between 1 and 43 hr. It is not to say that all of the smaller sample models have worse results, there are many examples with small errors, but the largest errors do tend to come from the less-popular models. We also find evidence of the more-popular models converging toward the same bias and uncertainty over time, as opposed to a lot more scatter between them in earlier times. Overall we conclude that the well-established models perform fairly similar with little bias and an uncertainty of order 10 – 15 hr and many of the less-established models may perform just as well but need more cases to further validate.

Data Availability Statement

The Arrival Time Scoreboard is available at <https://kauai.ccmc.gsfc.nasa.gov/CMEScoreboard/> and the DONKI catalog at <https://kauai.ccmc.gsfc.nasa.gov/DONKI/>. The specific dataset used in this work is archived via Zenodo at <https://doi.org/10.5281/zenodo.10932651> (Kay, 2024).

Acknowledgments

First and foremost we acknowledge the forecasters who submit their predictions to the ATSB. None of this work would have been possible without these existing sources or the teams at the CCMC and M2M Space Weather Analysis Office for populating and maintaining the online database. C. Kay would also like to thank E. Paouris for the useful discussion regarding the evolution of the metrics. C. Kay is supported by the National Aeronautics and Space Administration under Grant 80NSSC19K0909 issued through the Heliophysics Early Career Investigators program. E. Palmerio and P. Riley acknowledge support from NASA's Heliophysics Guest Investigators-Open program (Grant 80NSSC23K0447) as well as NSF's Prediction of and Resilience against Extreme Events program (Grant ICER-1854790).

References

- Alobaid, K. A., Abdullah, Y., Wang, J. T. L., Wang, H., Jiang, H., Xu, Y., et al. (2022). Predicting CME arrival time through data integration and ensemble learning. *Frontiers in Astronomy and Space Sciences*, *9*, 1013345. <https://doi.org/10.3389/fspas.2022.1013345>
- Amerstorfer, T., Möstl, C., Hess, P., Temmer, M., Mays, M. L., Reiss, M. A., et al. (2018). Ensemble prediction of a halo coronal mass ejection using heliospheric imagers. *Space Weather*, *16*(7), 784–801. <https://doi.org/10.1029/2017SW001786>
- Arge, C. N., Luhmann, J. G., Odstrcil, D., Schrijver, C. J., & Li, Y. (2004). Stream structure and coronal sources of the solar wind during the May 12th, 1997 CME. *Journal of Atmospheric and Solar-Terrestrial Physics*, *66*(15–16), 1295–1309. <https://doi.org/10.1016/j.jastp.2004.03.018>
- Arge, C. N., & Pizzo, V. J. (2000). Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates. *Journal of Geophysical Research*, *105*(A5), 10465–10480. <https://doi.org/10.1029/1999JA000262>
- Asvestari, E., Pomoell, J., Kilpua, E., Good, S., Chatzistergos, T., Temmer, M., et al. (2021). Modelling a multi-spacecraft coronal mass ejection encounter with EUFORIA. *Astronomy & Astrophysics*, *652*, A27. <https://doi.org/10.1051/0004-6361/202140315>
- Barnard, L., & Owens, M. (2022). HUXt—An open source, computationally efficient reduced-physics solar wind model, written in Python. *Frontiers in Physics*, *10*, 1005621. <https://doi.org/10.3389/fphy.2022.1005621>
- Bothmer, V., & Schwenn, R. (1998). The structure and origin of magnetic clouds in the solar wind. *Annales Geophysicae*, *16*, 1–24. <https://doi.org/10.1007/s00585-997-0001-x>
- Corona-Romero, P., Gonzalez-Esparza, J. A., Perez-Alanis, C. A., Aguilar-Rodriguez, E., de-la-Luz, V., & Mejia-Ambriz, J. C. (2017). Calculating travel times and arrival speeds of CMEs to Earth: An analytic tool for space weather forecasting. *Space Weather*, *15*(3), 464–483. <https://doi.org/10.1002/2016SW001489>
- Crosby, N. B., Veronig, A., Robbrecht, E., Vrsnak, B., Vennerstrom, S., Malandraki, O., et al. (2012). Forecasting the space weather impact: The COMESEP project. In Q. Hu, G. Li, G. P. Zank, X. Ao, O. Verkhoglyadova, & J. H. Adams (Eds.), *Space weather: The space radiation environment: 11th annual international astrophysics conference* (Vol. 1500, pp. 159–164). <https://doi.org/10.1063/1.4768760>
- Deforest, C., Killough, R., Gibson, S., Henry, A., Case, T., Beasley, M., et al. (2022). Polarimeter to UNify the corona and heliosphere (PUNCH): Science, status, and path to flight. In *2022 IEEE aerospace conference* (pp. 1–11). <https://doi.org/10.1109/AERO53065.2022.9843340>
- Dumbović, M., Čalogović, J., Vršnak, B., Temmer, M., Mays, M. L., Veronig, A., & Piantschitsch, I. (2018). The drag-based ensemble model (DBEM) for coronal mass ejection propagation. *The Astrophysical Journal*, *854*(2), 180. <https://doi.org/10.3847/1538-4357/aaa66>
- Feng, X., & Zhao, X. (2006). A new prediction method for the arrival time of interplanetary shocks. *Solar Physics*, *238*(1), 167–186. <https://doi.org/10.1007/s11207-006-0185-3>
- Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., & Vasyliunas, V. M. (1994). What is a geomagnetic storm? *Journal of Geophysical Research*, *99*(A4), 5771–5792. <https://doi.org/10.1029/93JA02867>
- Gopalswamy, N., Lara, A., Manoharan, P. K., & Howard, R. A. (2005). An empirical model to predict the 1-AU arrival of interplanetary shocks. *Advances in Space Research*, *36*(12), 2289–2294. <https://doi.org/10.1016/j.asr.2004.07.014>
- Gopalswamy, N., Lara, A., Yashiro, S., Kaiser, M. L., & Howard, R. A. (2001). Predicting the 1-AU arrival times of coronal mass ejections. *Journal of Geophysical Research*, *106*(A12), 29207–29218. <https://doi.org/10.1029/2001JA000177>
- Hess, P., & Zhang, J. (2015). Predicting CME ejecta and sheath front arrival at L1 with a data-constrained physical model. *The Astrophysical Journal*, *812*(2), 144. <https://doi.org/10.1088/0004-637X/812/2/144>
- Kay, C. (2024). Arrival time scoreboard 2013–2023. *Zenodo*. <https://doi.org/10.5281/ZENODO.10932651>
- Kay, C., Gopalswamy, N., Reinard, A., & Opher, M. (2017). Predicting the magnetic field of Earth-impacting CMEs. *The Astrophysical Journal*, *835*(2), 117. <https://doi.org/10.3847/1538-4357/835/2/117>
- Kay, C., Mays, M. L., & Collado-Vega, Y. M. (2022). OSPREI: A coupled approach to modeling CME-driven space weather with automatically generated, user-friendly outputs. *Space Weather*, *20*(4), e02914. <https://doi.org/10.1029/2021SW002914>

- Kay, C., Mays, M. L., & Verbeke, C. (2020). Identifying critical input parameters for improving drag-based CME arrival time predictions. *Space Weather*, 18(1), e2019SW002382. <https://doi.org/10.1029/2019SW002382>
- Kay, C., & Palmerio, E. (2024). Collection, collation, and comparison of 3D coronal CME reconstructions. *Space Weather*, 22(1), e2023SW003796. <https://doi.org/10.1029/2023SW003796>
- Kilpua, E. K. J., Balogh, A., von Steiger, R., & Liu, Y. D. (2017). Geoeffective properties of solar transients and stream interaction regions. *Space Science Reviews*, 212(3–4), 1271–1314. <https://doi.org/10.1007/s11214-017-0411-3>
- Kilpua, E. K. J., Lugaz, N., Mays, M. L., & Temmer, M. (2019). Forecasting the structure and orientation of earthbound coronal mass ejections. *Space Weather*, 17(4), 498–526. <https://doi.org/10.1029/2018SW001944>
- Kim, K. H., Moon, Y. J., & Cho, K. S. (2007). Prediction of the 1-AU arrival times of CME-associated interplanetary shocks: Evaluation of an empirical interplanetary shock propagation model. *Journal of Geophysical Research*, 112(A5), A05104. <https://doi.org/10.1029/2006JA011904>
- 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. *The Astrophysical Journal*, 855(2), 109. <https://doi.org/10.3847/1538-4357/aaae69>
- Lugaz, N., Temmer, M., Wang, Y., & Farrugia, C. J. (2017). The interaction of successive coronal mass ejections: A review. *Solar Physics*, 292(4), 64. <https://doi.org/10.1007/s11207-017-1091-6>
- Maharana, A., Scolini, C., Schmieder, B., & Poedts, S. (2023). Rotation and interaction of the CMEs of September 8 and 10, 2014, tested with EUHFORIA. *Astronomy & Astrophysics*, 675, A136. <https://doi.org/10.1051/0004-6361/202345902>
- Mayank, P., Vaidya, B., Mishra, W., & Chakrabarty, D. (2024). SWASTi-CME: A physics-based model to study coronal mass ejection evolution and its interaction with solar wind. *The Astrophysical Journal Supplement Series*, 270(1), 10. <https://doi.org/10.3847/1538-4365/ad08c7>
- Mays, M. L., Taktakishvili, A., Pulkkinen, A., MacNeice, P. J., Rastätter, L., Odstreil, D., et al. (2015). Ensemble modeling of CMEs using the WSA-ENLIL+Cone model. *Solar Physics*, 290(6), 1775–1814. <https://doi.org/10.1007/s11207-015-0692-1>
- McKenna-Lawlor, S. M. P., Dryer, M., Kartalev, M. D., Smith, Z., Fry, C. D., Sun, W., et al. (2006). Near real-time predictions of the arrival at Earth of flare-related shocks during Solar Cycle 23. *Journal of Geophysical Research (Space Physics)*, 111(A11), A11103. <https://doi.org/10.1029/2005JA011162>
- Möstl, C., Rollett, T., Frahm, R. A., Liu, Y. D., Long, D. M., Colaninno, R. C., et al. (2015). Strong coronal channelling and interplanetary evolution of a solar storm up to Earth and Mars. *Nature Communications*, 6(1), 7135. <https://doi.org/10.1038/ncomms8135>
- Núñez, M., Nieves-Chinchilla, T., & Pulkkinen, A. (2016). Prediction of shock arrival times from CME and flare data. *Space Weather*, 14(8), 544–562. <https://doi.org/10.1002/2016SW001361>
- Odstreil, D. (2003). Modeling 3-D solar wind structure. *Advances in Space Research*, 32(4), 497–506. [https://doi.org/10.1016/S0273-1177\(03\)00332-6](https://doi.org/10.1016/S0273-1177(03)00332-6)
- Odstreil, D. (2023). Heliospheric 3-D MHD ENLIL simulations of multi-CME and multi-spacecraft events. *Frontiers in Astronomy and Space Sciences*, 10, 1226992. <https://doi.org/10.3389/fspas.2023.1226992>
- Odstreil, D., & Pizzo, V. J. (1999a). Three-dimensional propagation of CMEs in a structured solar wind flow: 1. CME launched within the streamer belt. *Journal of Geophysical Research*, 104(A1), 483–492. <https://doi.org/10.1029/1998JA900019>
- Odstreil, D., & Pizzo, V. J. (1999b). Three-dimensional propagation of coronal mass ejections in a structured solar wind flow 2. CME launched adjacent to the streamer belt. *Journal of Geophysical Research*, 104, 493–504. <https://doi.org/10.1029/1998JA900038>
- Odstreil, D., Pizzo, V. J., Linker, J. A., Riley, P., Lionello, R., & Mikic, Z. (2004). Initial coupling of coronal and heliospheric numerical magnetohydrodynamic codes. *Journal of Atmospheric and Solar-Terrestrial Physics*, 66(15–16), 1311–1320. <https://doi.org/10.1016/j.jastp.2004.04.007>
- Palmerio, E., Kay, C., Al-Haddad, N., Lynch, B. J., Yu, W., Stevens, M. L., et al. (2021). Predicting the Magnetic Fields of a Stealth CME Detected by Parker Solar Probe at 0.5 au. *The Astrophysical Journal*, 920(2), 65. <https://doi.org/10.3847/1538-4357/ac25f4>
- Palmerio, E., Maharana, A., Lynch, B. J., Scolini, C., Good, S. W., Pomoell, J., et al. (2023). Modeling a Coronal Mass Ejection from an Extended Filament Channel. II. Interplanetary Propagation to 1 au. *The Astrophysical Journal*, 958(1), 91. <https://doi.org/10.3847/1538-4357/ad0229>
- Palomba, M., & Luntama, J.-P. (2022). Vigil: ESA space weather mission in L5. In *44th cospar scientific assembly. Held 16-24 July* (Vol. 44).3544.
- Paouris, E., & Mavromichalaki, H. (2017). Effective acceleration model for the arrival time of interplanetary shocks driven by coronal mass ejections. *Solar Physics*, 292(12), 180. <https://doi.org/10.1007/s11207-017-1212-2>
- Paouris, E., & Vourlidas, A. (2022). Time-of-Arrival of coronal mass ejections: A two-phase kinematics approach based on heliospheric imaging observations. *Space Weather*, 20(7), e2022SW003070. <https://doi.org/10.1029/2022SW003070>
- Pizzo, V., Millward, G., Parsons, A., Biesecker, D., Hill, S., & Odstreil, D. (2011). Wang-sheeley-arge-enlil cone model transitions to operations. *Space Weather*, 9(3), 03004. <https://doi.org/10.1029/2011SW000663>
- Pomoell, J., & Poedts, S. (2018). EUHFORIA: European heliospheric forecasting information asset. *Journal of Space Weather and Space Climate*, 8, A35. <https://doi.org/10.1051/swsc/2018020>
- Reiss, M. A., Muglach, K., Mullinix, R., Kuznetsova, M. M., Wiegand, C., Temmer, M., et al. (2023). Unifying the validation of ambient solar wind models. *Advances in Space Research*, 72(12), 5275–5286. <https://doi.org/10.1016/j.asr.2022.05.026>
- Richardson, I. G. (2018). Solar wind stream interaction regions throughout the heliosphere. *Living Reviews in Solar Physics*, 15(1), 1. <https://doi.org/10.1007/s41116-017-0011-z>
- Riley, P., & Ben-Nun, M. (2022). sunRunner1D: A tool for exploring ICME evolution through the inner heliosphere. *Universe*, 8(9), 447. <https://doi.org/10.3390/universe8090447>
- Riley, P., Mays, M. L., Andries, J., Amerstorfer, T., Biesecker, D., Delouille, V., 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, J. A., Kubicka, M., Amerstorfer, U. V., & Harrison, R. A. (2016). EIEvoHI: A novel CME prediction tool for heliospheric imaging combining an elliptical front with drag-based model fitting. *The Astrophysical Journal*, 824(2), 131. <https://doi.org/10.3847/0004-637X/824/2/131>
- Rotter, T., Veronig, A. M., Temmer, M., & Vršnak, B. (2015). Real-time solar wind prediction based on SDO/AIA coronal hole data. *Solar Physics*, 290(5), 1355–1370. <https://doi.org/10.1007/s11207-015-0680-5>
- Schwenn, R., dal Lago, A., Huttunen, E., & Gonzalez, W. D. (2005). The association of coronal mass ejections with their effects near the Earth. *Annales Geophysicae*, 23(3), 1033–1059. <https://doi.org/10.5194/angeo-23-1033-2005>
- Scolini, C., Chané, E., Temmer, M., Kilpua, E. K. J., Dissauer, K., Veronig, A. M., et al. (2020). CME-CME interactions as sources of CME geoeffectiveness: The formation of the complex ejecta and intense geomagnetic storm in 2017 early September. *The Astrophysical Journal - Supplement Series*, 247(1), 21. <https://doi.org/10.3847/1538-4365/ab6216>

- Scolini, C., Rodriguez, L., Mierla, M., Pomoell, J., & Poedts, S. (2019). Observation-based modelling of magnetised coronal mass ejections with EUHFORIA. *Astronomy & Astrophysics*, 626, A122. <https://doi.org/10.1051/0004-6361/201935053>
- Shi, T., Wang, Y., Wan, L., Cheng, X., Ding, M., & Zhang, J. (2015). Predicting the arrival time of coronal mass ejections with the graduated cylindrical shell and drag force model. *The Astrophysical Journal*, 806(2), 271. <https://doi.org/10.1088/0004-637X/806/2/271>
- Singh, T., Benson, B., Raza, S. A. Z., Kim, T. K., Pogorelov, N. V., Smith, W. P., & Arge, C. N. (2023). Improving the arrival time estimates of coronal mass ejections by using magnetohydrodynamic ensemble modeling, heliospheric imager data, and machine learning. *The Astrophysical Journal*, 948(2), 78. <https://doi.org/10.3847/1538-4357/acc10a>
- Smith, Z. K., Dryer, M., McKenna-Lawlor, S. M. P., Fry, C. D., Deehr, C. S., & Sun, W. (2009). Operational validation of HAFv2's predictions of interplanetary shock arrivals at Earth: Declining phase of Solar Cycle 23. *Journal of Geophysical Research (Space Physics)*, 114(A5), A05106. <https://doi.org/10.1029/2008JA013836>
- Tobiska, W. K., Knipp, D., Burke, W. J., Bouwer, D., Bailey, J., Odstrcil, D., et al. (2013). The Anemomilos prediction methodology for Dst. *Space Weather*, 11(9), 490–508. <https://doi.org/10.1002/swe.20094>
- Vandas, M., Fischer, S., Dryer, M., Smith, Z., & Detman, T. (1996). Parametric study of loop-like magnetic cloud propagation. *Journal of Geophysical Research*, 101(A7), 15645–15652. <https://doi.org/10.1029/96JA00511>
- Verbeke, C., Mays, M. L., Kay, C., Riley, P., Palmerio, E., Dumbović, M., et al. (2023). Quantifying errors in 3D CME parameters derived from synthetic data using white-light reconstruction techniques. *Advances in Space Research*, 72(12), 5243–5262. <https://doi.org/10.1016/j.asr.2022.08.056>
- Vourlidis, A., Patsourakos, S., & Savani, N. P. (2019). Predicting the geoeffective properties of coronal mass ejections: Current status, open issues and path forward. *Philosophical Transactions of the Royal Society of London, Series A*, 377(2148), 20180096. <https://doi.org/10.1098/rsta.2018.0096>
- Vršnak, B., Žic, T., Vrbanec, D., Temmer, M., Rollett, T., Möstl, C., et al. (2013). Propagation of interplanetary coronal mass ejections: The drag-based model. *Solar Physics*, 285(1–2), 295–315. <https://doi.org/10.1007/s11207-012-0035-4>
- Wang, J., Ao, X., Wang, Y., Wang, C., Cai, Y., Luo, B., et al. (2018). An operational solar wind prediction system transitioning fundamental science to operations. *Journal of Space Weather and Space Climate*, 8, A39. <https://doi.org/10.1051/swsc/2018025>
- Wang, Y., Liu, J., Jiang, Y., & Erdélyi, R. (2019). CME arrival time prediction using convolutional neural network. *The Astrophysical Journal*, 881(1), 15. <https://doi.org/10.3847/1538-4357/ab2b3e>
- Wold, A. M., Mays, M. L., Taktakishvili, A., Jian, L. K., Odstrcil, D., & MacNeice, P. (2018). Verification of real-time WSA-ENLIL+Cone simulations of CME arrival-time at the CCMC from 2010 to 2016. *Journal of Space Weather and Space Climate*, 8(27), A17. <https://doi.org/10.1051/swsc/2018005>
- Wu, C.-C., Dryer, M., Wu, S. T., Wood, B. E., Fry, C. D., Liou, K., & Plunkett, S. (2011). Global three-dimensional simulation of the interplanetary evolution of the observed geoeffective coronal mass ejection during the epoch 1–4 August 2010. *Journal of Geophysical Research (Space Physics)*, 116(A12), A12103. <https://doi.org/10.1029/2011JA016947>
- Xie, H., Ofman, L., & Lawrence, G. (2004). Cone model for halo CMEs: Application to space weather forecasting. *Journal of Geophysical Research (Space Physics)*, 109(A3), 3109. <https://doi.org/10.1029/2003JA010226>
- Zhao, X., & Dryer, M. (2014). Current status of CME/shock arrival time prediction. *Space Weather*, 12(7), 448–469. <https://doi.org/10.1002/2014SW001060>
- Zhao, X. H., & Feng, X. S. (2014). Shock Propagation Model version 2 and its application in predicting the arrivals at Earth of interplanetary shocks during Solar Cycle 23. *Journal of Geophysical Research (Space Physics)*, 119(1), 1–10. <https://doi.org/10.1002/2012JA018503>
- Zhao, X. P., Plunkett, S. P., & Liu, W. (2002). Determination of geometrical and kinematical properties of halo coronal mass ejections using the cone model. *Journal of Geophysical Research*, 107(A8), 1223. <https://doi.org/10.1029/2001JA009143>
- Zheng, Y., & Rastaetter, L. (2015). Space weather products and tools used in auroral monitoring and forecasting at CCMC/SWRC. *Geophysical Monograph Series*, 215, 291–301. <https://doi.org/10.1002/9781118978719.ch19>