Advancing Space Weather Prediction: Machine Learning and Bayesian Modelling for CMEs and Coronal Jets

Simone Chierichini

Submitted for the degree of Doctor of Philosophy

School of Mathematical and Physical Sciences

March 2025

Supervisor: Prof. Robertus Erdélyi, Prof. Dario Del Moro

University of Sheffield

PhD in Astronomy, Astrophysics and Space Science

PhD Cycle XXXVII

Advancing Space Weather Prediction: Machine Learning and Bayesian Modelling for CMEs and Coronal Jets

Simone Chierichini

Supervisors: Prof. Dario Del Moro
Prof. Robertus Erdélyi

Coordinator: Prof. Francesco Piacentini, Prof. Giuseppe Bono

Academic Year 2023/2024

Contents

List of Figures

List of Tables

List of Acronyms

SW	Solar Wind
DBM	Drag Based Model
P-DBM	Probabilistic Drag Based Model
CME	Coronal Mass Ejection
ICME	Interplanetary Coronal Mass Ejection
SDO	Solar Dynamics Observatory
AIA	Atmospheric Imaging Assembly
SOHO	Solar and Heliospheric Observatory
ACE	Advanced Composition Explorer
MCMC	Monte Carlo Markov Chain
EUV	Extreme Ultraviolet
ML	Machine Learning
AI	Artificial Intelligence
SVM	Support Vector Machine
SVR	Support Vector Regressor
ViT	Vision Transformer
TNT	Transformer in Transformer
CNN	Convolutional Neural Network
CAT-PUMA	CME Arrival Time Prediction Using Machine learning Algorithms
GSE	Geocentric Solar Ecliptic System
TRACE	Transition Region and Coronal Explorer
SOHO	Solar and Heliospheric Observatory
LASCO	Large Angle and Spectrometric Coronagraph
NASA	National Aeronautics and Space Administration
MPA	Measurement Position Angle
ANOVA	Analysis of Variance
MAE	Mean Absolute Error
MSE	Mean Squared Error
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
FPR	False Positive Rate
TPR	True Positive Rate

FAR	False Alarm Ratio
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
CV	Cross Validation
L1	Lagrangian-1
MHD	Magnetohydrodynamic
WSA	Wang-Sheeley-Arge
SHAP	Shapley Additive exPlanation
EUHFORIA	EUropean Heliospheric FORecasting Information Asset
SAJIA	Semi-Automated Jet Identification Algorithm
MM	Mathematical Morphology
ToA	time of arrival
VoA	velocity of arrival
DL	Deep Learning
PDF	probability density function
BSV	Best Split Validation
ACF	auto-correlation function
PSRF	Potential Scale Reduction Factor
SE	structuring element
TPE	Tree-structured Parzen Estimator
GPS	Global Positioning System
RBF	Radial basis function
KNN	k-nearest neighbors
MPA	Measurement Position Angle
DBEM	Drag based Ensemble model
TPU	tensor processing unit
GPU	graphic processing unit
MI	mutual information
GIF	Graphics Interchange Format
AdaBoost	adaptive boosting
XGBoost	Extreme gradient boosting
DDPMs	Denoising Diffusion Probabilistic Models
HD	Hydrodinamc
IRIS	Interface Region Imaging Spectrograph
PSP	Parker Solar Probe

Abstract

This thesis investigates the use of Machine Learning (ML) and Bayesian inference to improve the prediction and understanding of Coronal Mass Ejection (CME), a critical aspect of space weather forecasting.

Several ML techniques, including supervised learning methods such as support vector machines, decision trees, and ensemble methods, are used to develop predictive models based on CME data, aiming to enhance the accuracy of CME arrival time forecasts. A key focus is placed on model interpretability, achieved through Shapley Additive exPlanation (SHAP) values, which provide insights into the feature space and allow for a better understanding of how different variables influence model outputs.

Additionally, the thesis applies Bayesian inference and Monte Carlo Markov Chain (MCMC) techniques to refine probabilistic models of CME propagation using drag-based models, further improving the robustness and reliability of the predictions.

The work also extends ML applications to the study of other solar phenomena, specifically coronal jets, by augmenting the dataset for jet identification. This leads to increased dataset diversity, improved detection of rare events, and a better understanding of solar dynamics.

Overall, this thesis presents advancements in the application of ML and Bayesian techniques to space weather forecasting and the study of solar phenomena. The tools and methods developed in this research hold considerable potential for future applications, with the capacity to improve prediction accuracy and mitigate the impacts of space weather on technological systems.

Chapter 1

Introduction

The Sun, our closest stellar neighbour, is a sphere of hot plasma situated at the centre of the Solar System. Its gravitational force governs the orbits of all celestial bodies within the Solar System, including planets, asteroids, and comets [Schrijver and Siscoe, 2010]. Predominantly composed of hydrogen (approximately 74%) and helium (about 24%), with trace amounts of heavier elements, the Sun is classified as a G-type main-sequence star (G2V) and has been emitting energy for about 4.6 billion years [Bahcall et al., 2000]. The Sun's internal structure comprises several distinct layers, each contributing uniquely to its overall function:

The core, the innermost region, is the site of nuclear fusion. At temperatures reaching approximately 15 million K, hydrogen nuclei fuse to form helium, releasing vast amounts of energy [Clayton, 1984, Kravvaris et al., 2023]. Surrounding the core, the radiative zone extends to about 70% of the Sun's radius. Here, energy is transported outward through radiative diffusion, with photons scattering off particles and gradually moving towards the outer layers. Beyond the radiative zone lies the convective zone, where the temperature is sufficiently low for convection currents to develop. Hot plasma rises towards the surface, cools, and then sinks back down to be reheated, creating convective motion that efficiently transfers energy to the Sun's surface [Nordlund et al., 2009].

The photosphere, the Sun's visible surface, emits the light what we see. It is relatively cooler compared to the inner layers, with temperatures around 5,500 K [Stix, 1989]. Situated above the photosphere, the chromosphere appears as a reddish rim during solar eclipses. Temperatures in this layer range from 4,000 to 25,000 K, and it is the site of dynamic phenomena such as solar flares and prominences [Murawski et al., 2020]. Beyond the chromosphere lies the Sun's outermost atmospheric layer, the corona.

It is the outermost layer of the Sun's atmosphere, extending millions of kilometres into space. Despite its greater distance from the core, the corona is much hotter than the underlying layers, with temperatures reaching several million of K. It can be best observed during a total solar eclipse or with specialised instruments and is the source of the interplanetary medium, the solar wind [Aschwanden, 2005].

The Sun's dynamic processes, collectively known as solar activity, have a profound impact on the Earth and the entire Solar System. The Sun's activity has significant implications for space weather, which can impact

Earth's technological systems and space environment [Thaduri et al., 2020, Buzulukova and Tsurutani, 2022].

1.1 Solar activity

Observations of the Sun's activity reveal a near-periodic cycle, typically close to a period of 11 years, commonly referred to as the solar cycle [Withbroe, 1989, Hathaway, 2010]. One of the primary indicators used to monitor solar activity is the number of sunspots on the photosphere. Sunspots are dark areas on the solar surface characterized by strong, constantly shifting magnetic fields. By convention, a solar activity cycle begins with a minimum number of sunspots and ends with the onset of the next minimum. During a minimum period, the number of sunspots is relatively low, while a maximum period is associated with a significantly higher number of sunspots. Moreover, the position of sunspots changes throughout the cycle; they are typically distributed near the equator during minimum periods and migrate towards intermediate latitudes as the maximum approaches [Solanki et al., 2006]. The sunspot record provides a direct means of characterizing solar activity over nearly 400 years [Hathaway and Wilson, 2004], showing that sunspot cycles have periods of 131 ± 14 months with a normal distribution, are asymmetric with a fast rise and slow decline, and that the rise time from minimum to maximum decreases with cycle amplitude [Hathaway and Wilson, 2004]). The longest recorded period was 17.1 years (from 1788 to 1805), while the shortest lasted 7.3 years (from 1829 to 1837). The solar cycle is closely linked to variations in the Sun's magnetic field, which reverses its polarity approximately every 11 years, causing the North and South Poles to switch places [Charbonneau, 2010].

This cyclic behaviour indicates the existence of an internal timing mechanism within the Sun that appears to control or influence all aspects of solar phenomena and extends its effects throughout the Solar System. It is widely accepted that this timing mechanism results from the nearly periodic generation and evolution of magnetic fields within the solar interior and on the solar surface, specifically through a dynamo mechanism that generates the Sun's magnetic field [Charbonneau, 2010, Balogh et al., 2014]. The cyclic nature of solar activity is evident in most parameters used to describe solar phenomena, even during extended intervals when sunspots were scarce. Indeed, even in the absence of sunspots, other indicators of solar activity persist, including emissions of ultraviolet and X-ray radiation, the modulation of the interplanetary medium, the solar wind, and the occurrence of energetic events such as solar flares and coronal mass ejections. These manifestations of solar activity can have a significant impact on the Earth's upper atmosphere, ionosphere, and magnetosphere, collectively known as "space weather," and may also influence certain aspects of the lower atmosphere and climate [Baker, 2000].

In the past few decades, the understanding of the solar activity cycle has shifted from being a scientific interest to a matter of practical importance. This change is driven by the realisation that solar phenomena can have potentially

harmful effects on Earth's space environment. Solar activity associated with space weather can be divided into three main components: solar flares, CMEs, and high-speed solar wind [Schrijver and Siscoe, 2010, Gopalswamy, 2018].

Solar flares are large, sudden eruptions of intense electromagnetic radiation from the Sun, lasting from minutes to hours. These powerful bursts of energy occur due to the rapid reconfiguration and reconnection of the Sun's magnetic field lines, releasing vast amounts of stored magnetic energy [Hudson, 1991]. The electromagnetic radiation, traveling at the speed of light, interacts with Earth's upper atmosphere as soon as the flare is observed, causing immediate ionospheric disturbances on the sunlit side. These disturbances can disrupt satellite operations and degrade radio communications, highlighting the significant impact of solar flares on space weather and technological systems [Baker et al., 2013].

High-speed solar wind streams originate from areas on the Sun known as coronal holes. These magnetically open regions of the Sun's atmosphere allow the accelerated outflow of high-speed solar wind particles, which can reach velocities of up to 800 kilometres per second. These coronal holes can form at any latitude on the Sun, but their resulting solar wind streams usually only impact the Earth when they are located nearer to the solar equator. This is because equatorial coronal holes are more favourably oriented to interact with and influence the Earth's magnetosphere and upper atmosphere, leading to more pronounced space weather effects [McPherron and Weygand, 2006, Cranmer, 2009, Al-Feadh and Al-Ramdhan, 2019].

CMEs are immense expulsions of plasma and magnetic fields from the Sun's outer atmosphere, or corona [Gopalswamy et al., 2006]. These massive eruptions can propel billions of tonnes of solar material into interplanetary space at millions of kilometres per hour. CMEs can be ejected in any direction relative to the Earth, but only those aimed towards our planet will have significant impacts. When a CME reaches and interacts with the Earth's magnetosphere, the compressed and distorted magnetic fields and influx of charged solar particles can disrupt satellite operations, communication systems, and power grids, demonstrating the substantial effects of these solar phenomena on space weather and terrestrial technology [Gopalswamy et al., 2006].

1.2 Space weather

Recent decades have seen a growing acknowledgement of the significant influence that space-based phenomena can have on human activities and endeavours on Earth [Camporeale et al., 2018c]. This enhanced awareness highlights the crucial importance of comprehending and forecasting space weather in order to mitigate its impacts on contemporary technology and human welfare [Lanzerotti, 2001, Schwenn, 2006, Pulkkinen, 2007, Temmer, 2021, Cliver et al., 2022].

Space weather phenomena originate primarily from the Sun, where complex magnetic fields are generated in the outer "convective" layer. In this region, hot plasma rises towards the solar surface, transporting energy

and distorting the magnetic fields. As the Sun's differential rotation further twists and contorts these fields, they emerge into the Sun's outer atmosphere - the corona - forming intricate magnetic structures visible in extreme ultraviolet imaging. The energy stored within these magnetic fields is the driving force behind space weather, particularly through the process of magnetic reconnection. This process reconfigures the magnetic field topology in the solar corona, releasing vast amounts of energy and accelerating electrons and ions in the plasma to extremely high velocities [Singh et al., 2010]. This reconfiguration can cause portions of the corona to become magnetically disconnected from the Sun, allowing them to be ejected into interplanetary space as powerful bursts of charged particles and electromagnetic radiation (known as CMEs) [Camporeale et al., 2018b]. Space weather events primarily impact three key environments, which can significantly disrupt the operation of many critical technologies that are essential for modern societies and economies:

- The electromagnetic fields within the solid body of the Earth can induce disruptive currents in power grids, pipelines, and other ground-based infrastructure.
- The radiation environments in Earth's atmosphere and near-Earth space can degrade the performance and lifetime of satellites, endanger astronauts, and disturb radio communications and navigation signals.
- The density, composition, and dynamics of the upper atmosphere can affect the orbital trajectories of satellites, compromise the integrity of Global Positioning System (GPS) signals, and disrupt high-frequency radio communications.

Disturbances caused by space weather can profoundly impact a wide range of essential systems and infrastructure. These effects extend beyond the immediate disruption of space-based telecommunications, broadcasting, weather services, and navigation. They also significantly disrupt power distribution networks and terrestrial communications, particularly at higher latitudes where the impacts are more pronounced. A notable consequence of solar activity is the severe disruption of satellite navigation services, which is caused by dynamic changes in the ionosphere. This poses significant challenges for aviation, road transport, shipping, and any other activities that rely on precise positioning and timing information provided by satellite navigation. The far-reaching implications of these space weather-induced disruptions are substantial, compromising the safety and efficiency of numerous critical operations that modern societies and economies depend upon [Fry, 2012, Rao et al., 2009].

CMEs are considered the most significant phenomena through which solar activity drives space weather effects on Earth. CMEs are vast eruptions of magnetised plasma from the Sun's outer atmosphere that propagate through interplanetary space. When a fast-moving CME reaches Earth, it can profoundly interact with and compress the planet's magnetosphere, causing a sudden and substantial increase in the magnetic field observed

at the Earth's surface. This dynamic interaction between the CME and the magnetosphere is a key driver of intense geomagnetic storms, which are the primary manifestations of severe space weather. Geomagnetic storms can induce powerful electrical currents in the upper atmosphere and on the ground, posing significant risks to power grids, communication networks, satellite operations, and other critical technological systems that modern societies rely upon. Understanding the initiation and evolution of CMEs, as well as their complex interactions with the Earth's magnetic field, is, therefore, a crucial aspect of space weather research and forecasting efforts. The interaction between fast-moving CMEs and Earth's magnetosphere often involves a complex process of magnetic reconnection. This allows the CME's magnetic field to directly connect with and effectively "plug into" Earth's own magnetic field, enabling a substantial transfer of energy from the CME into the magnetosphere. This sudden influx of energy can drive a dynamic cycle of energy storage and explosive release within the magnetosphere's tail region. The stored energy is eventually unleashed in the form of powerful electrical currents, which surge back towards Earth, producing vibrant auroras, heating the upper atmosphere, and generating strong electric currents that flow through the ionosphere. This "substorm cycle" is a fundamental yet intricate dynamic process inherent to the interaction between planetary magnetospheres and space weather events with impact. For Earth, this substorm cycle typically unfolds over the course of 1–3 hours [Durgonics et al., 2017, Chakraborty et al., 2020].

The passage of a large, fast-moving CME through Earth's interplanetary space is not an instantaneous event. Observations indicate that CME durations can vary widely, ranging from as short as 2 hours to as long as 90 hours, with an average duration of about 20–23 hours [Gopalswamy, 2006, Richardson and Cane, 2010]. In the case of particularly fast CMEs, the full extent of the event may sweep past our planet within 12–24 hours, depending on the velocity and structural characteristics of the CME [Gopalswamy, 2006]. During this time, the CME's interaction with the magnetosphere drives a series of recurring substorms, resulting in a prolonged period of heightened space weather effects known as a geomagnetic storm [Badruddin et al., 2018]. These geomagnetic storms can have significant and wide-ranging impacts on both ground-based and space-borne technological systems. From power grids and pipelines to satellite operations and communication networks, the disruptive effects of geomagnetic storms pose serious challenges to the critical infrastructure that modern societies and economies rely upon. Given the potential severity of space weather effects, there has been substantial effort over the past few decades directed towards forecasting if and when CMEs will arrive at Earth. Accurate prediction of these events can provide critical lead time to mitigate their impacts. These space weather events primarily originate from the Sun, particularly from the evolution of its magnetic field. However, solar activity is not constant; it follows a cyclic process during which the configuration of magnetic field lines undergoes modifications. Understanding these cycles and their implications for space weather is crucial for developing effective forecasting models. Efforts to forecast CME arrival and their subsequent

impacts involve a combination of observational data, theoretical models, and, increasingly, machine learning techniques. By integrating data from solar observations with advanced modelling approaches, researchers aim to improve the accuracy and reliability of space weather predictions, thereby enhancing our ability to protect technological infrastructure and human activities from the adverse effects of space weather.

1.3 Coronal mass ejections

CMEs are massive, high-energy structures composed of plasma and magnetic fields, ejected from the Sun's corona into interplanetary space. These eruptions typically have an average mass of approximately 10^{13} kg and plasma temperatures ranging from 80,000 K to 2 million K. The frequency of CMEs is closely tied to the solar cycle, with the Sun producing up to 2–3 CMEs per day during solar maximum due to enhanced magnetic activity and instability in the corona.

The first recorded observation of a solar eruption dates back to the 1859 Carrington Event, when Richard Carrington observed a solar flare accompanied by a CME [Carrington, 1860]. Magnetic disturbances caused by the CME were later confirmed through ground-based magnetometer readings. Early observations of transient solar events were limited to rare total solar eclipses due to the Sun's intense glare. However, advancements in space-based solar observatories have revolutionized the study of CMEs, providing continuous and detailed monitoring of these dynamic phenomena.

CMEs are the most massive eruptive events in the solar system, ejecting magnetized plasma clouds at velocities of millions of kilometers per hour. These structures, often spanning millions of kilometers, originate in closed magnetic field regions of the corona and are frequently associated with filament or prominence eruptions, which accompany around 70% of CMEs. Solar flares often occur alongside CMEs, with simultaneous events observed in 55–90% of cases, depending on the energy released [Green et al., 2002, Youssef, 2012].

The discovery of CMEs was made in the early 1970s through space-borne coronagraphs aboard missions like Skylab. Subsequent advancements with observatories such as the Solar Maximum Mission, Yohkoh, Solar and Heliospheric Observatory (SOHO), and Transition Region and Coronal Explorer (TRACE) have provided extensive data, improving our understanding of their morphology and behavior.

The energy driving CMEs primarily comes from free magnetic energy stored in the solar atmosphere's non-potential magnetic fields, with pressure and gravitational forces also contributing. However, measuring the coronal magnetic field remains challenging, requiring reliance on extrapolated photospheric measurements [Mikić and Lee, 2006, Kusano et al., 2012].

A comprehensive CME structure typically includes the following components [Green et al., 2018b]:

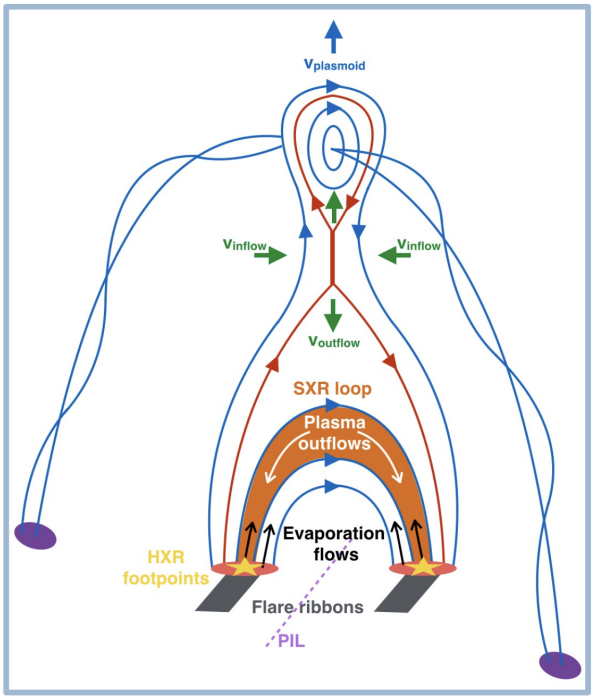


FIGURE 1.1: An illustrative drawing of the CSHKP Model [Carmichael, 1964, Sturrock, 1966, Hirayama, 1974, Kopp and Pneuman, 1976].

- A fast-moving shock wave ahead of the CME, compressing and heating the ambient solar wind plasma.
- A leading edge with elevated plasma density ($n_e \approx 10^{14}\,\mathrm{m}^{-3}$) and magnetic field intensity ($10^{-4}\,\mathrm{T}$).
- A cavity with reduced plasma density ($n_e \approx 10^{13}\,\mathrm{m}^{-3}$) and coronal temperatures of 1–2 MK.
- A prominence core with high plasma density ($n_e \approx 10^{17}\,\mathrm{m}^{-3}$) and lower temperatures ($\sim 80,000\,\mathrm{K}$).
- A post-eruption arcade with plasma temperatures of 10 MK.

The widely accepted CSHKP model [Carmichael, 1964, Sturrock, 1966, Hirayama, 1974, Kopp and Pneuman, 1976] explains CME initiation through magnetic reconnection. An unstable pre-eruptive structure rises, stretching magnetic field lines until reconnection occurs. This process ejects plasma into interplanetary space while forming post-eruption loops, as illustrated in Fig. 1.1 [Priest and Forbes, 2001].

The evolution of a CME/flare event typically occurs in three phases [Green et al., 2018a]: initiation, characterized by the gradual rise of pre-eruptive structures [Forbes, 2000, Mittal and Narain, 2010]; acceleration, marked by the violent ejection of plasma; and propagation, during which the ejected material travels at nearly constant velocity unless influenced by solar wind interactions [Vršnak, 2008].

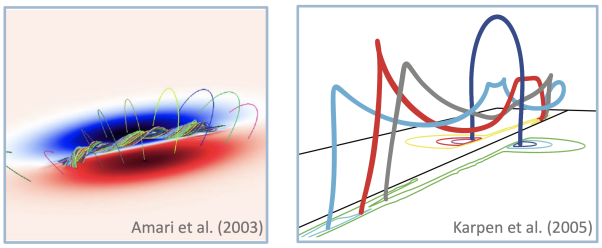


FIGURE 1.2: Pre-eruptive magnetic field configurations. Flux rope structure (left)[Amari et al., 2003]. Sheared arcade (right) [Karpen et al., 2005].

Two main pre-eruptive configurations, flux ropes and sheared arcades, underpin CMEs. Recent studies suggest that these may form a hybrid state, evolving from magnetic loop emergence to flux rope structures, depending on the dynamics of the solar region [Wang et al., 2015, Zheng et al., 2020, Patsourakos et al., 2020], as shown in Fig. 1.2.

1.3.1 CME propagation and effects

As mentioned in the previous section, a CME typically follows three distinct evolutionary phases:

- **The initiation phase**: The frontal loop rises slowly (at $\sim$ 80 km/s) as the CME is triggered.
- **The acceleration phase**: The frontal loop undergoes rapid acceleration, lasting from several to tens of minutes, often coinciding with the impulsive phase of an associated flare.
- **The propagation phase**: The frontal loop moves at an approximately constant velocity.

The velocity of a CME is generally measured as the radial propagation speed of its frontal loop, projected in the plane of the sky[1].

Sheeley Jr et al. [1999] classified CMEs into two types based on height-time maps observed by SOHO/Large Angle and Spectrometric Coronagraph (LASCO) coronagraph:

- **Gradual CMEs**: Formed as prominences and their cavities rise beneath coronal streamers, with speeds ranging between 400–600 km/s.
- **Impulsive CMEs**: Typically associated with flares, with speeds exceeding 750 km/s.

Statistical studies (e.g., Zhang and Dere [2006]) show that the main acceleration during the impulsive phase varies widely, from 2.8×10^{-3} to 4.464 km/s, with an average value of 0.331 km/s. The duration of this phase ranges from 6 to 1200 minutes, averaging around 180 minutes.

[1]This velocity is often referred to as projected velocity.

Although the propagation phase is characterised by a near-constant speed, smaller accelerations or decelerations can occur. Fast CMEs tend to decelerate, while slow CMEs accelerate, such that their velocities approach the ambient solar wind speed.

It is common to classify CMEs as "fast" or "slow" based on their initial speed relative to the ambient solar wind and their resulting acceleration or deceleration. Fast CMEs (initial speed well above solar wind speed) tend to decelerate, and slow CMEs (below solar wind speed) tend to accelerate, as the drag force drives them toward the ambient flow speed [MacQueen and Fisher, 1983, Sheeley Jr et al., 1999, Gopalswamy et al., 2000, Vršnak et al., 2001]. However, this binary classification is not universally accepted. Observational studies suggest a continuum of CME kinematics rather than a clear bimodal separation [Pant et al., 2021]. In other words, CME speeds and accelerations span a broad range without a distinct gap between "slow" and "fast" populations, and intermediate-speed CMEs show mixed kinematic behavior. The apparent categories likely overlap, and the drag-driven kinematic profile of each CME depends on a continuum of parameters (e.g. launch speed, mass, and ambient conditions) rather than falling neatly into two discrete groups [Pant et al., 2021]. It is therefore more accurate to treat CME propagation speeds as a spectrum, which has important implications for modelling—each event may need individualised treatment instead of assuming one of two standard kinematic profiles.

The propagation is influenced by the highly variable solar wind, the presence of preceding CMEs (common during solar maximum), and interactions with other CMEs or nearby coronal holes. Intrinsic driving properties within the CME itself may also play a role [Chen, 2011, Webb and Howard, 2012].

CMEs often deflect and rotate as they travel outward, due to interactions with structured background magnetic fields and solar wind flows. Near the Sun, influences such as active region fields, coronal hole open flux, and helmet streamers can push a CME away from a purely radial trajectory and even alter its orientation (rotation) [Isavnin et al., 2013, Kay et al., 2015, Cécere et al., 2023]. Further out, interactions with high-speed solar wind streams can continue to deflect a CME's path or rotate its flux-rope axis, as observed in cases where a CME's encounter with a fast stream caused it to change direction and tilt in interplanetary space [Palmerio et al., 2022]. CMEs also undergo magnetic erosion during propagation: magnetic reconnection with the surrounding interplanetary field strips away outer layers of the CME's magnetic flux and mass [Dasso et al., 2006, Ruffenach et al., 2015]. This erosion effectively slows the CME's forward motion (by reducing its momentum and cross-sectional area), and can delay its arrival at Earth by several hours in the case of fast CMEs [Stamkos et al., 2023]. Each of these processes – deflection, rotation, and erosion – is crucial for space-weather forecasting, as they alter the CME's expected arrival time and impact. A CME that deflects or rotates might miss an intended target or deliver a different magnetic orientation than expected, and erosion can diminish a CME's magnetic intensity while also affecting its transit time [Palmerio et al., 2022, Stamkos et al., 2023].

The dynamics of CME evolution in interplanetary space remain incompletely understood, but several forecasting methods have been developed to predict their arrival times at Earth [Napoletano et al., 2018b, Brueckner et al., 1998, Owens et al., 2005, Vršnak et al., 2014]. These methods include:

- **Experimental models**: Empirical models relying on observational statistical relationships between coronal measurements and heliospheric propagation parameters.
- **Magnetohydrodynamic (MHD)-based models**: Numerical simulations of Interplanetary Coronal Mass Ejection (ICME) propagation, which are computationally intensive and require detailed knowledge of the heliospheric state.
- **Hybrid models**: Simplified analytical or empirical approaches based on MHD or Hydrodinamc (HD) frameworks, requiring modest computational power to describe interactions during CME propagation.

1.4 Solar wind

Once launched into interplanetary space, CMEs propagates through the solar wind, a continuous outward flow of plasma from the Sun's corona. The hypothesis of the solar wind was first proposed in the mid-20th century to explain the behaviour of comet tails, which often deviated from the expected Sun-comet vector. These deviations suggested the existence of a radial flow of charged particles from the Sun. This theory, which postulated the continuous expulsion of plasma primarily composed of electrons and protons at velocities of 500–1000 km/s, was first proposed by Biermann [1952].

The solar corona, a dynamic magnetic environment, consists of loops and structures anchored in the Sun's photosphere. Most of these loops are closed and trap plasma, but in regions with weaker or open magnetic fields, plasma can flow outward along field lines, forming the solar wind [Gosling, 2006, undefined, 2020].

In 1958, Eugene Parker introduced the concept of the solar wind through a theoretical model that explained this continuous plasma outflow, even before it was directly observed [Parker, 1958]. Parker's groundbreaking model provided a unified explanation for various phenomena, including comet tail behaviour. The Soviet Luna 1 spacecraft in 1959 and National Aeronautics and Space Administration (NASA)'s Mariner 2 mission in 1962 provided the first direct observational evidence, validating Parker's predictions [Sonett, 1963].

Parker's model assumes a steady plasma flow from a spherically symmetric Sun with an isothermal corona. He demonstrated that the corona's temperature and pressure are sufficiently high to overcome the Sun's gravitational pull, resulting in a continuous outward plasma flow. The governing equations for momentum and mass conservation in the model are:

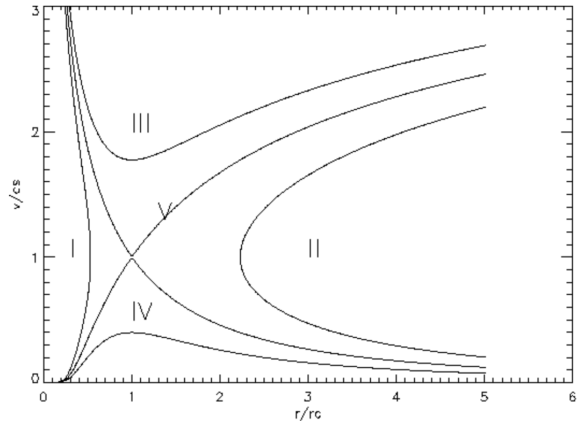


FIGURE 1.3: Solar wind velocity v as a function of radial distance, illustrating the five solution classes of Parker's motion equations [Parker, 1958].

$$\rho(\boldsymbol{u}\cdot\nabla)\boldsymbol{u} = -\nabla P + \boldsymbol{j}\times\boldsymbol{B} + \rho\boldsymbol{F}_g, \tag{1.1}$$

$$\nabla\cdot(\rho\boldsymbol{u}) = 0, \tag{1.2}$$

where $\boldsymbol{u}$ is the radial expansion speed, ρ is the density, and the forces on the right-hand side represent pressure gradients, the Lorenz force, and gravitational force, respectively. Combining these equations and the model's assumptions yields the equation of motion:

$$\left(u^2 - \frac{2k_BT}{m}\right)\frac{1}{u}\frac{du}{dr} = \frac{4k_BT}{mr} - \frac{GM_\odot}{r^2}, \tag{1.3}$$

where u is the solar wind's radial velocity as a function of distance r, k_B is the Boltzmann constant, T is the corona's temperature, m is the particle mass (typically a proton), G is the gravitational constant, and $M_\odot$ is the Sun's mass.

This equation balances the forces acting on solar wind particles: the outward thermal pressure and the inward gravitational pull. Parker showed that as coronal temperatures rise, the thermal energy surpasses gravitational constraints, driving the solar wind's expansion. Of the five possible solutions to this equation (Fig. 1.3), only the solar wind solution (Class V) aligns with observations, exhibiting subsonic velocities near the Sun and supersonic speeds beyond the critical point.

The solar wind is generally categorized into two types based on speed: the *fast* solar wind and the *slow* solar wind. Fast solar wind originates in coronal holes and propagates at speeds exceeding 650 km/s, while slow solar wind, associated with streamer belts, has velocities below 400 km/s.

1.4.1 Fast and slow solar wind

Observations have shown that solar wind properties differ markedly depending on velocity, often categorized into "fast" and "slow" regimes. Here,

high-speed wind is conventionally considered to have velocities around $\geq 650\,\mathrm{km/s}$, featuring elevated temperatures, reduced densities, and lower mass fluxes indicative of a more tenuous plasma outflow. Conversely, low-speed wind is often described at velocities of $\leq 400\,\mathrm{km/s}$, characterized by cooler temperatures, higher densities, and higher mass fluxes, suggesting a denser and somewhat less energetic flow [Bravo and Stewart, 1997]. Speeds falling between these thresholds are typically viewed as transitional or intermediate regimes, exhibiting properties that blend aspects of both slow and fast wind [Zhao et al., 2024, Alterman et al., 2024].

Fast solar wind streams typically originate from active regions such as polar and equatorial coronal holes, which are areas of open magnetic field lines that allow the rapid outflow of plasma at velocities around 700-800 km/s. These high-speed streams are often referred to as Parker's "classical" solar wind.

In contrast, the slow solar wind has a more controversial origin, as it can arise from a variety of complex and dynamic solar structures. The slow wind can emanate from the edges of polar coronal holes, where the magnetic field is less tightly concentrated, as well as from small low-latitude coronal hole streamer belts. Additionally, the slow solar wind may even originate from the fringes of active regions [Ohmi et al., 2004], where the magnetic field topology is more open and conducive to a more gradual plasma outflow at velocities around 300-400 km/s.

The structure of the solar wind gives rise to a phenomenon known as the "Parker spiral", named after Eugene Parker. This is due to the "frozen-in" condition, where the magnetic field is "frozen" into the solar wind flow, as described by Alfvén's pioneering work [Alfvén, 1942]. As the solar wind emanates radially outward from the Sun's surface, the rotation of the Sun causes the footpoints of the magnetic field lines to move faster than the outer portions of the magnetic field, which are carried along by the outflowing plasma. This differential rotation between the Sun's surface and the outer solar wind results in the magnetic field lines being stretched into a spiral pattern, forming the characteristic Parker spiral structure. In essence, it is the plasma that "pulls" the magnetic field along as it propagates through the interplanetary medium, imparting this distinctive spiral configuration to the overall solar wind structure. The differing speeds of the fast and slow solar wind can lead to complex interactions between the two wind types. This is due to the distinct velocity regimes that characterise each component of the solar wind. When a region producing slow solar wind is followed by a region generating fast solar wind, the faster wind can catch up to and interact with the slower wind. This interaction generates a rarefaction region, where the plasma density is lower, and a compression region, where the density is higher. The boundary between the fast and slow wind regions is known as the stream interface, which acts as a separator between the two wind types. This interface can also generate reflection waves in the solar wind, creating a distinctive structure composed of the interface itself, bracketed by the reflection waves. This structure is known as a corotating interaction region, reflecting the role of the Sun's rotation in shaping the solar wind dynamics [Rouillard et al., 2008].

1.5 Drag-based model

The Drag Based Model (DBM) offers a simplified representation of CME propagation dynamics in the solar wind, drawing an analogy to aerodynamic drag exerted by the interplanetary medium [Owens and Grandé, 2004].

This model assumes that beyond a certain distance from the Sun, interplanetary CMEs tend to adjust their velocity to match the ambient solar wind, with faster ICMEs decelerating and slower ones accelerating. While this is consistent with the idea that at larger distances, the solar wind's drag force becomes the dominant factor in ICME kinematics, the process is not straightforward, as the magnetic field intensity and upstream solar wind conditions also influence the deceleration and acceleration of ICMEs. [Gopalswamy et al., 2000].

In this framework, the radial acceleration of a CME is determined by the solar wind speed $w(r)$ and the drag parameter $\gamma(r)$, following the equation:

$$a = -\gamma(r)[(v - w(r))]|v - w(r)|, \tag{1.4}$$

where v represents the CME velocity, and r is the distance from the Sun. The drag parameter γ encapsulates information about the interaction between ICMEs and the solar wind and can be expressed as a function of the ICME cross-sectional area (A), the solar wind density (ρ_w), the ICME mass (M), and the virtual mass ($M_v \approx \rho_w \frac{V}{2}$, where V is the ICME volume) [Cargill, 2004]. It is typically expressed as:

$$\gamma = \frac{c_d A \rho_w}{M_v + M}, \tag{1.5}$$

where c_d is the drag coefficient. In general, the drag parameter γ may vary with time, but it is reasonable to assume that $\gamma(r)$ and $w(r)$ remain constant beyond approximately 20 solar radii [Cargill, 2004, Vršnak et al., 2013]. Under this assumption, one can obtain the CME velocity $v(t)$ and the heliospheric distance $r(t)$ as functions of time:

$$v(t) = \frac{v_0 - w}{1 \pm \gamma(v_0 - w)t} + w, \tag{1.6}$$

$$r(t) = \pm\frac{1}{\gamma}\ln\left[1 \pm \gamma(v_0 - w)t\right] + wt + r_0, \tag{1.7}$$

where v_0 represents the initial CME velocity, and r_0 is the initial heliospheric distance.

The DBM framework allows us to make predictions for the time of arrival (ToA) and the impact velocity (velocity of arrival (VoA)) of a CME by fixing the travelled distance ($r_1 - r_0 \approx 1$ AU) and using the DBM parameters as inputs. In Napoletano et al. [2018a], a probabilistic version of the DBM was introduced, referred to as Probabilistic Drag Based Model (P-DBM), which employs *a-priori* distributions of γ and w to obtain estimates of ToA and VoA along with their associated errors.

Classical DBM assume a steady aerodynamic drag force that slows

fast CMEs and speeds up slow ones, using a drag coefficient and a drag force proportional to the relative speed between the CME and solar wind. Recent refinements introduce turbulent drag-force models, which account for high–Reynolds number conditions in the solar wind by making the drag force scale non-linearly (quadratically) with the CME–solar wind speed difference [Cranmer et al., 2021]. This approach, inspired by turbulence, effectively increases drag on fast CMEs and has been shown in simulations to better reproduce CME deceleration profiles than the linear drag law of the classical DBM [Subramanian et al., 2012, Lin and Chen, 2022]. By incorporating turbulent flow effects, these refined models improve the physical realism of CME propagation forecasts without sacrificing the simplicity that makes DBM useful.

1.6 Coronal jets

Solar activity encompasses a diverse range of dynamic and energetic phenomena that originate from the Sun's magnetic field. Among these, coronal jets stand out as distinct, collimated bursts of plasma that erupt from the solar corona, often associated with magnetic reconnection events [Shibata et al., 1992, Raouafi et al., 2016, Liu et al., 2023]. These jets are highly dynamic and variable, showcasing the Sun's remarkable ability to channel its magnetic energy into powerful and structured flows [Török et al., 2015, Raouafi et al., 2016]. Their eruptive nature situates them as key components within the broader framework of solar activity, serving as vital linkages to other prominent phenomena such as solar flares and CMEs.

Coronal jets occupy a unique place in solar physics as small-scale but profoundly revealing phenomena. First observed in the 1970s, their discovery was tied to advancements in solar observations, which began unveiling transient eruptions in the corona. Early studies speculated that these jets could contribute to heating the corona and accelerating the solar wind, making them relevant to key questions in heliophysics. Over time, space missions like Yohkoh, SOHO, Hinode and Interface Region Imaging Spectrograph (IRIS) have provided a wealth of data, showing that coronal jets are far more than isolated events. Instead, they bridge the gap between small-scale transients like spicules and large-scale eruptions such as CMEs. These jets reveal the importance of magnetic energy and dynamics of the Sun, acting as a lens to understand its broader activity.

Coronal jets are observed across multiple wavelengths, notably in Extreme Ultraviolet (EUV) and X-ray spectra, allowing for detailed analysis of their thermal and dynamic properties [Nisticò et al., 2009]. They typically last from a few minutes to tens of minutes and can reach lengths of several tens of thousands of kilometers. The velocities of these jets vary, with some reaching speeds up to several hundred kilometers per second [Raouafi et al., 2016]. High-resolution observations from instruments such as the Hinode, Solar Dynamics Observatory (SDO) and IRIS satellites have provided valuable data on the fine-scale structure and evolution of coronal jets [Young and Muglach, 2014, Raouafi et al., 2016, Schmieder et al., 2022]. These studies

have revealed the presence of helical structures and rotational motions within jets, suggesting the involvement of twisted magnetic fields and the release of magnetic helicity during the reconnection process [Liu et al., 2017, Chen et al., 2020].

Observations have identified various types of coronal jets, which differ in their underlying physical mechanisms and characteristics. "Standard" jets exhibit a well-defined, narrow spire, reflecting a more localised magnetic reconnection process. In contrast, "blowout" jets are more expansive and involve broader eruptions, indicative of a more complex and dynamic magnetic field configuration [Moore et al., 2010b, Morton et al., 2012, Li et al., 2018]. This distinction is linked to the nature of the magnetic reconnection occurring in these regions, with blowout jets suggesting a more dramatic release of energy and rapid reconfiguration of the magnetic field [Raouafi et al., 2016]. The specific characteristics of these jet types provide valuable insights into the diverse range of magnetic field topologies and energy release processes operating in the solar corona.

1.6.1 Role in solar activity

Coronal jets are not isolated phenomena, but rather interconnected with broader solar dynamics and activities. They play a crucial role in the mass and energy balance of the solar corona, and are considered potential drivers of the solar wind, particularly the fast solar wind emanating from coronal holes. The frequent occurrence of these jets in these specific regions supports the hypothesis that they actively contribute to the acceleration of solar wind particles, providing a crucial link between solar surface phenomena and the heliosphere beyond Török et al. [2015], Raouafi et al. [2016], Lionello et al. [2016], Chitta et al. [2023].

Similarly to major dynamic solar events such as solar flares and CMEs, theoretical and observational studies suggest that coronal jets are primarily driven by magnetic reconnection [Shibata et al., 1992, Canfield et al., 1996, Moore et al., 2010a, Pariat et al., 2015, Sterling et al., 2015]. By analyzing these dynamic features, researchers can gain valuable insights into the specific conditions that trigger reconnection, the intricate mechanisms governing the rate at which it occurs, and the broader implications for energy transfer within the solar atmosphere. However, alternative MHD models propose a distinct mechanism for jet formation, one driven not by magnetic flux emergence, but by the injection of helicity through photospheric motions [Pariat et al., 2015, 2016, Raouafi et al., 2016]. Specifically, shear or twisting motions at the base of a closed non-potential region beneath a preexisting null point can induce magnetic reconnection with surrounding quasi-potential flux, resulting in untwisting or helical jets [see, e.g. Pariat et al., 2015]. Furthermore, previous studies have shown that the evolution of coronal jets is often preceded by wave-like or oscillatory disturbances [Pucci et al., 2012, Scullion et al., 2012, Bagashvili et al., 2018].

Observations have revealed that many jets are associated with oscillations in coronal emissions near the jet bases, likely driven by changes in the area

or temperature of the pre-jet region [Pucci et al., 2012]. Statistically, pre-jet intensity oscillations have been observed approximately 12-15 minutes before the onset of jets [Bagashvili et al., 2018], which may be linked to the generation of MHD waves arising from rapid temperature variations and shear flows during local reconnection events [Shergelashvili et al., 2006].

Furthermore, their chromospheric counterparts, spicules (with a length of approximately 10 Mm), play a key role in coronal heating and solar wind acceleration [De Pontieu et al., 2004, Shibata et al., 2007, Tian et al., 2014, Dey et al., 2022, Liu et al., 2023, Kesri et al., 2024], although the exact mechanisms remain unclear.

Recent research is revealing connections between small-scale solar eruptions and the magnetic switchbacks detected by Parker Solar Probe (PSP) in the young solar wind [Bale et al., 2019]. Switchbacks are sudden reversals in the magnetic field direction, and one emerging idea is that they originate from interchange magnetic reconnection in the low corona – the same process that drives coronal jets [Sterling and Moore, 2020, Drake et al., 2021]. Coronal jets are produced when closed loop magnetic fields reconnect with adjacent open flux, releasing plasma and untwisting field into the heliosphere. It has been proposed that the kinked or S-shaped field structures generated by these jet-producing reconnection events travel outward and evolve into switchbacks in the solar wind [Sterling and Moore, 2020]. Evidence for this link is growing: recent studies have mapped switchback-rich solar wind streams back to their solar sources and found correlations with jet activity at coronal hole boundaries, where interchange reconnection is active. In particular, a statistical analysis by Hou et al. [2024] found a direct relationship between the magnetic flux changes from reconnection-driven polar jets and the magnetic deflections observed as switchbacks, suggesting a physically intrinsic connection between coronal jets and switchback formation. These findings support the view that interchange reconnection events on the Sun (jet eruptions, jetlets, etc.) inject perturbations that manifest as Alfvénic switchback structures in the solar wind [Sterling and Moore, 2020, Hou et al., 2024]. This emerging link is important because it connects phenomena across scales — from transient jets at the Sun to in-situ switchback measurements — improving our understanding of solar wind magnetic structure and the role of small-scale solar activity in shaping heliospheric conditions.

Coronal jets play a crucial role in the complex interplay between the Sun's magnetic field and its outer atmosphere, providing important insights into the fundamental processes that drive space weather Török et al. [2015], Lionello et al. [2016]. By studying these dynamic features, researchers can better understand the intricate mechanisms that govern the Sun's activity and its far-reaching impacts on the Earth and the broader heliosphere.

1.7 Problem statement and objectives

The accurate forecasting of CMEs is a critical challenge in space weather research due to their significant impact on Earth's technological systems. These massive plasma and magnetic field eruptions from the Sun's corona

interact with the solar wind and Earth's magnetosphere, often resulting in geomagnetic storms that can disrupt satellites, power grids, and communication networks. While forecasting methods have advanced significantly, limitations in data, model interpretability, and computational efficiency persist, necessitating further exploration.

Researchers have developed various approaches to predict the arrival and impact of CMEs. Among these, three primary approaches have gained prominence, each leveraging distinct methodologies and offering unique advantages and limitations.

MHD models are one of the foundational approaches to simulating CMEs propagation through the heliosphere. These models solve complex plasma equations, capturing detailed interactions between CMEs and the surrounding solar wind. Well-established models such as Wang-Sheeley-Arge (WSA)-ENLIL and EUropean Heliospheric FORecasting Information Asset (EUHFORIA) [Odstrcil, 2003, Pomoell and Poedts, 2018] have significantly contributed to our understanding of CME dynamics and their potential impacts on Earth's space environment.

Significant advances in 3D CME modelling have improved our ability to predict CME arrival times and impact [Isavnin, 2016, Maharana et al., 2022, Scolini and Palmerio, 2024]. Modern heliospheric models now simulate CMEs with fully three-dimensional structures (e.g. as flux-rope volumes or deformable spheroids) within ambient solar wind flows, in contrast to earlier 1D or 2D formulations that often treated the CME as a point or circle. These 3D models capture the CME's expansion, distortion, and interaction with solar wind structures more realistically, leading to better forecasts. For example, the introduction of a 3D flux-rope CME model (FRi3D) into the EUHFORIA simulation framework has enabled more accurate modeling of CME flank interactions and magnetic field profiles at Earth, substantially improving arrival predictions over simpler cone or spheromak models [Isavnin, 2016, Maharana et al., 2022]. The ability to include the CME's true geometry and orientation in simulations means that effects like non-radial propagation, pancaking (front flattening), and shear interactions with high-speed streams are now accounted for, reducing errors in transit time estimates. Overall, these 3D and physics-rich models (including full MHD simulations and ensemble modeling in three dimensions) are narrowing the gap between predicted and observed CME arrival times and enhancing the reliability of space weather forecasts.

Despite their effectiveness, MHD models are computationally demanding, requiring significant time and advanced computational resources. This high computational cost often limits their feasibility for real-time forecasting, where rapid predictions are critical during space weather events.

While these approaches demonstrated reasonably effective prediction of CME arrival times, they often require significant computational resources. Furthermore, a class of HD-based models has emerged, which rely on the hypothesis that the dynamics of CMEs in interplanetary space are solely governed by their interaction with the surrounding solar wind [Cargill, 2004, Owens and Cargill, 2004, Shi et al., 2015]. One popular model in this

category is the DBM [Vršnak et al., 2013, Cargill, 2004, Napoletano et al., 2018a, Dumbović et al., 2018, Mugatwala et al., 2024, Chierichini et al., 2024a].

Despite substantial efforts, the accuracy of prediction times of arrival remains constrained by limitations in available data. These limitations stem from the challenges of characterizing CME properties at launch using remote sensing observations and the inability to accurately model the inner heliosphere.

In a prior study, Napoletano et al. [2018a] introduced a probabilistic version of the DBM, termed the P-DBM, to address the dearth of information and provide estimates of the inherent uncertainty in CME forecasts. The P-DBM approach replaces the constant DBM parameters with a-priori probability distributions, leveraging an ensemble modelling framework to generate probability density functions of ToA and VoA at a target location. This framework enables the production of the most probable ToA and VoA estimates along with their associated prediction uncertainty [e.g. Del Moro et al., 2019, Piersanti et al., 2020]. In a subsequent study, Napoletano et al. [2022] proposed a modified version of these probability density functions (PDFs) employing an inversion procedure of DBM equations based on a Monte Carlo-like approach. Napoletano et al. [2018a] and Napoletano et al. [2022] explore the possibility that the probability distribution functions of the DBM parameters may vary depending on the type of solar wind accompanying the propagation of CMEs. The dynamics of CMEs are modelled as that of a solid body moving in a fluid stream, suggesting that an appropriate description of the propagation dynamics is required for accelerated or decelerated CMEs. On average, they improved the understanding of the parameter PDFs, leading to enhanced prediction of the arrival time.

Additionally, ML represents a growing alternative to traditional physical and semi-empirical methods, leveraging data-driven techniques to uncover patterns and relationships in historical observations. Unlike the predefined physical frameworks of MHD or semi-empirical models, ML approaches adaptively learn from the data, making them well-suited to handle non-linear and high-dimensional dependencies [Camporeale et al., 2018a]. For instance, the CME Arrival Time Prediction Using Machine learning Algorithms (CAT-PUMA) model [Liu et al., 2018] uses CME and solar wind parameters at launch to predict CME transit times with a mean absolute error (Mean Absolute Error (MAE)) of 5.9 hours. Deep Learning (DL) methods, such as Convolutional Neural Networks (CNNs), extend this capability by processing raw observational data, including white-light images, to forecast CME arrival times [Wang et al., 2019, Fu et al., 2021]. These techniques offer computational efficiency and the potential for real-time application. However, their success depends on the availability and quality of training datasets, which remains a challenge given the relatively small sample of Earth-impacting CMEs. Moreover, the interpretability of some ML models can be limited, which may pose challenges for stakeholders requiring insights into the underlying physics. Despite these obstacles, the rapid advancement of ML techniques and their ability to integrate diverse datasets make them a promising tool for the future of CME forecasting.

Turning to the broader question of coronal jets, one may ask whether there is a solar cycle effect on these localized dynamic features. While solar cyclic activity is characterized by global phenomena such as the long-term evolution of sunspot numbers, solar irradiance variations, and the frequency of solar flares and CMEs [Solanki and Krivova, 2011, Song et al., 2016, Bhowmik and Nandy, 2018], the extent to which the solar cycle influences smaller-scale solar features like coronal jets remains poorly understood, underscoring the need for comprehensive statistical studies. Shimojo et al. [1996] conducted an early study of 100 jets, primarily originating from active regions, over the period from November 1991 to April 1992, using a manual examination of X-ray observations from the Yohkoh Soft X-Ray Telescope [Ogawara, 1995].

Bennett and Erdélyi [2015] investigate macrospicules using high-resolution observations from the SDO, focusing on their spatial and temporal properties. By examining their statistical characteristics, the research establishes relationships between velocity, length, and lifetime, providing a foundation for theoretical modeling. Additionally, a correlation between macrospicule properties and the solar activity cycle is identified, suggesting an influence of solar minimum-to-maximum transitions.

A comprehensive statistical analysis of 301 macrospicules cunducted by Kiss et al. [2017] over 5.5 years is conducted using SDO/Atmospheric Imaging Assembly (AIA) data at 30.4 nm. The study examines variations in macrospicule properties across different solar regions and hemispheres, revealing periodic oscillations with a cycle just under two years. Furthermore, a pronounced hemispheric asymmetry in macrospicule distribution suggests a link between global solar dynamo processes and local atmospheric phenomena.

Gyenge et al. [2017] explore the connection between active longitudes and CME occurrences, using morphological parameters such as sunspot tilt angle and separateness. Findings indicate that the most complex active regions, associated with increased magnetic helicity and fast CMEs, cluster around active longitudes. The study highlights the potential of active longitude-based forecasting for CME sources and provides insights for solar dynamo modeling.

Moreover, [Kiss and Erdélyi, 2018] examine the influence of the global solar magnetic field on macrospicules through a seven-year observational dataset. A wavelet analysis of macrospicule properties reveals periodicities around two years (quasi-biennial oscillations). A comparison with solar activity proxies exhibiting similar oscillations shows an out-of-phase relationship, suggesting that global solar activity may modulate local chromospheric dynamics.

More recently, Liu et al. [2023] introduced the Semi-Automated Jet Identification Algorithm (SAJIA), which was applied to data from the SDO/AIA [Lemen et al., 2012] during solar cycle 24, from 2010 to 2020. This study identified 1215 jets and revealed power-law distributions between intensity/energy and frequency, along with quasi-annual oscillations, offering new insights into the temporal behavior of these solar phenomena.

Recent advancements in machine learning have further uncovered cyclic patterns in solar activity. For instance, Diercke et al. [2024] applied DL to detect

a *filament cycle* based on H-alpha observations during Solar Cycle 24, while Zhang et al. [2024] revealed a *prominence cycle* using a similar approach with SDO/AIA 304 Å images. These findings underscore the increasing application of machine learning techniques in identifying cyclical solar phenomena. Such advancements are crucial for understanding the connection between large-scale solar cycles and localized phenomena such as coronal jets, which is vital for improving space weather forecasting and its implications for Earth's technological systems.

In another recent study, Bourgeois et al. [2025] utilized SDO/AIA images to investigate long-term properties of coronal off-limb structures throughout Solar Cycle 24, employing Mathematical Morphology (MM), a technique that focuses on the analysis of geometric structures. This method enabled a detailed examination of both eruptive and atmospheric solar phenomena, yielding comprehensive statistics and critical insights into active longitudes, thus enhancing our understanding of solar dynamics during the cycle.

MM, a powerful tool for image enhancement, shape analysis, and feature detection, has gained prominence in solar physics for identifying and tracking solar events like filaments [Shih and Kowalski, 2003, Koch and Rosolowsky, 2015, Barata et al., 2018, Carvalho et al., 2020, Bourgeois et al., 2024]. Although MM originated in the 1960s [Matheron, 1967, Haas et al., 1967, Serra, 1969], its widespread application in solar research is relatively recent. Combining SAJIA and MM allows us to create a more comprehensive dataset, leveraging the strengths of both methods to offer a richer foundation for studying and predicting solar jet phenomena.

This thesis focuses on advancing semi-empirical and ML-based methods for CME forecasting, addressing limitations in data availability, interpretability, and uncertainty quantification. Additionally, it applies ML techniques to enhance datasets of coronal jets, smaller-scale solar eruptions that contribute to our understanding of solar activity and space weather dynamics.

The research is structured into three interconnected projects:

1.7.1 CME arrival modelling with machine learning

Building on Liu et al. [2018], this project investigates the CAT-PUMA model, which predicts CME transit times using CME launch characteristics and solar wind features. While the model demonstrates promising accuracy, its performance has only been evaluated on limited datasets. This project expands the dataset to include CME events up to 2022, enabling a more comprehensive evaluation and refinement of CAT-PUMA. A variant of the model is also developed to classify Earth-impacting CMEs, providing actionable insights for mitigating geomagnetic storm risks. Additionally, interpretability techniques are applied to uncover the physical relationships captured by the model.

1.7.2 Improving the P-DBM with bayesian inference

Semi-empirical models like the DBM benefit from computational efficiency but lack robust mechanisms for quantifying uncertainty. The P-DBM [Napoletano

et al., 2018a] addresses this by introducing probabilistic parameter distributions. Building on this foundation, this project employs Markov Chain Monte Carlo (MCMC) methods to refine these distributions and improve the accuracy and reliability of CME arrival time forecasts. By incorporating solar wind conditions into the parameter estimation process, the P-DBM provides a more nuanced understanding of CME propagation dynamics, particularly under fast and slow solar wind regimes [Napoletano et al., 2022].

1.7.3 Augmenting coronal jet datasets with machine learning and mathematical morphology

Coronal jets, narrow and transient eruptions within the solar corona, share characteristics with CMEs and play a role in space weather dynamics. While prior studies, such as Liu et al. [2023], have identified coronal jets using the SAJIA algorithm, this thesis integrates SAJIA with MM techniques [Bourgeois et al., 2025] to enhance the dataset. MM, a powerful image processing tool, allows for the identification and characterization of coronal structures. By combining these methods, the work presented in this thesis produces a richer dataset of jets, enabling detailed statistical analyses of their spatial and temporal distributions and potential connections to the solar cycle.

This thesis seeks to address critical challenges in CME forecasting and coronal jet analysis by employing advanced statistical and machine learning techniques. Through improvements in model accuracy, uncertainty quantification, and dataset enrichment, the research makes significant strides toward enhancing our ability to predict and mitigate the impacts of space weather, safeguarding technological systems in an increasingly interconnected world.

This thesis is structured as follows:

Chapter 2 details the methodological framework employed throughout the thesis. It describes the application of machine learning techniques, including supervised learning methods like Random Forests and Support Vector Machines, and Bayesian inference approaches, particularly the use of MCMC techniques for parameter estimation. Additionally, the chapter discusses MM methods used to enhance datasets related to coronal jets. A particular emphasis is placed on model interpretability, with tools like SHAP values used to provide insights into the significance of various features in the models.

In Chapter 3, the sources and preprocessing of the datasets used in the study are discussed. This includes observational data of CMEs from instruments such as SOHO LASCO and in-situ solar wind measurements from Advanced Composition Explorer (ACE), as well as data related to coronal jets. The chapter describes steps taken to clean, augment, and engineer features from these datasets to improve their utility for modeling. The challenges of working with sparse and noisy solar data are also addressed, with a focus on how these limitations were mitigated in the research.

Chapter 4 presents the key findings of the thesis. For CMEs, the integration of machine learning and Bayesian inference methods resulted in improved transit time predictions, with the inclusion of probabilistic techniques enhancing uncertainty quantification. In the case of coronal jets, the application of random forest algorithms and dataset augmentation techniques expanded the available catalog of events, providing a more diverse and robust dataset for studying solar dynamics. The chapter compares the models' performance against prior studies and provides a detailed analysis of key metrics, as well as an interpretation of feature importance rankings.

Finally, Chapter 5 is devoted to interpreting the results in the broader context of space weather research. The discussion emphasizes the implications of the improved forecasting models for operational space weather prediction and highlights the strengths and limitations of the employed methods. The potential for future advancements is also explored. The thesis closes by reflecting on its contribution to the field, particularly in advancing the understanding and forecasting of solar phenomena.

The Appendix A provides supplementary materials that support the main content of the thesis. This include extended background on Bayesian theory.

Chapter 2

Methods

This chapter provides a comprehensive overview of the theoretical foundations underpinning the methodologies employed throughout this thesis. It begins with an introduction to machine learning, highlighting key algorithms and techniques used for predictive modeling and classification tasks. Bayesian inference is then discussed, focusing on its application for probabilistic modeling and parameter estimation, with a particular emphasis on MCMC methods, which play a crucial role in quantifying uncertainties and exploring complex parameter spaces. Finally, the chapter delves into mathematical morphology, a powerful tool for analyzing and processing geometrical structures in solar observations. Together, these methods form the backbone of the analytical approaches used to investigate and model solar phenomena in this thesis.

2.1 Machine learning

ML has emerged as a transformative technology, revolutionizing diverse domains such as finance, healthcare, transportation, and entertainment [Sen et al., 2021, Sarker, 2021, Roy et al., 2023]. Unlike traditional programming, where developers must explicitly define every step of a process, ML enables computers to autonomously learn from data, uncover complex patterns, and make accurate predictions or informed decisions. This paradigm shift has allowed computers to address problems previously thought to require human intelligence and cognition [LeCun et al., 2015, Sejnowski, 2020].

The rapid progress in ML has been fueled by the increasing availability of digital data, advancements in computational power, and the development of sophisticated algorithms. The vast influx of data from sources such as social media and sensors provides a rich resource for training ML models. Simultaneously, specialized hardware like graphic processing units (GPUs) and tensor processing units (TPUs) has enabled efficient training and deployment of complex models [Thompson et al., 2020].

Rooted in the interdisciplinary fields of statistics and artificial intelligence, ML has evolved significantly over the past decades [Naqa and Murphy, 2015]. Advances in computational power, access to diverse datasets, and algorithmic innovations have led to a versatile range of ML techniques tailored to address various types of problems and data structures.

Key ML approaches include supervised learning, where models are trained on labeled data to map inputs to outputs; unsupervised learning, which discovers patterns and structures within unlabeled data [Bishop and Nasrabadi, 2006, Srinivasa et al., 2018]; semi-supervised learning, which combines labeled and unlabeled data to improve performance [van Engelen and Hoos, 2019]; and reinforcement learning, where agents learn optimal decision-making through trial and error, guided by rewards and penalties [Li, 2017, François-Lavet et al., 2018]. These methodologies have enabled the creation of intelligent systems capable of tasks such as speech recognition, image processing, predictive analytics, and autonomous decision-making.

2.1.1 Supervised learning

Supervised learning is a widely employed machine learning approach that trains a model on labeled data, where input features $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ are paired with corresponding target outputs $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$. The model learns to map $\mathbf{X}$ to $\mathbf{Y}$ by minimizing the error between its predictions $\hat{\mathbf{Y}}$ and the actual outcomes $\mathbf{Y}$. This optimization process is represented as:

$$\min_{\theta} \mathcal{L}(\mathbf{Y}, f(\mathbf{X}; \theta)),$$

where $\mathcal{L}$ is the loss function, f represents the model with parameters θ, and $\hat{\mathbf{Y}} = f(\mathbf{X}; \theta)$ are the predictions. Supervised learning is particularly effective for classification tasks, where models predict discrete class labels by maximizing probabilities, and regression tasks, where models predict continuous outputs by minimizing errors such as the mean squared error (Mean Squared Error (MSE)):

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

2.1.2 Linear regression

Linear regression, a fundamental supervised learning method, models the relationship between $\mathbf{X}$ and $\mathbf{Y}$ as a linear function:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m,$$

or, in vectorized form:

$$\hat{y} = \mathbf{X}\boldsymbol{\theta},$$

where $\mathbf{X}$ is the feature matrix and $\boldsymbol{\theta}$ the parameter vector, including the intercept θ_0. The goal is to find $\boldsymbol{\theta}$ that minimizes the sum of squared errors between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$:

$$\min_{\boldsymbol{\theta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}).$$

The optimal $\boldsymbol{\theta}$ can be determined analytically using the normal equation:

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Alternatively, iterative methods such as gradient descent are commonly used. Gradient descent updates the parameters iteratively to minimize the loss function:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}),$$

where α is the learning rate and $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})$ is the gradient of the loss function. Repeatedly applying this rule ensures convergence to the optimal parameters.

While linear regression is effective for modeling linear relationships, many real-world problems involve complex, non-linear patterns. To capture these, advanced machine learning models, such as neural networks and ensemble methods, extend the capabilities of supervised learning to tackle a broader range of applications.

Support vector machines

Support Vector Machine (SVM) are versatile supervised learning methods widely used for classification, regression, and outlier detection [Cortes and Vapnik, 1995]. They are grounded in statistical learning theory and excel in tasks requiring robust generalization, particularly in high-dimensional spaces. The key idea behind SVMs is to construct an optimal hyperplane that separates data into distinct classes while maximizing the margin, defined as the distance between the hyperplane and the closest data points, known as support vectors.

Given a training dataset $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^d$ are feature vectors and $y_i \in \{-1, 1\}$ are class labels, a linear SVM aims to find a hyperplane defined by the weight vector $\mathbf{w}$ and bias b that satisfies:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2,$$

subject to:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad \forall i.$$

This optimization problem is solved using quadratic programming techniques to identify the hyperplane with maximum margin, ensuring better generalization performance. For non-linear decision boundaries, SVMs leverage kernel functions $K(\mathbf{x}_i, \mathbf{x}_j)$, enabling implicit transformations into higher-dimensional feature spaces without explicitly computing the coordinates [Schölkopf et al., 1999]. Common kernels include:

- Polynomial kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^p$, - Radial Basis Function (Radial basis function (RBF)) kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, - Sigmoid kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\alpha\mathbf{x}_i \cdot \mathbf{x}_j + c)$.

The dual formulation introduces Lagrange multipliers α_i, transforming the optimization problem into:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j),$$

subject to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C,$$

where C controls the trade-off between maximizing margin and minimizing classification error.

SVMs are also adaptable to datasets with overlapping classes through soft margin SVMs, which introduce slack variables ξ_i to allow some misclassifications:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i,$$

subject to:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

Here, C balances margin maximization and misclassification penalties, making SVMs robust to noise and overlapping data. For regression tasks, Support Vector Regressor (SVR) employs a similar approach, aiming to fit a hyperplane within a tolerance margin ϵ while minimizing prediction errors [Smola and Schölkopf, 2004]:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \max(0, |y_i - f(X_i)| - \epsilon).$$

SVMs have been successfully applied across domains such as image recognition, bioinformatics, and text categorization due to their robustness against overfitting and flexibility in handling non-linear relationships through appropriate kernels [Ben-Hur et al., 2008, Awad and Khanna, 2015].

Decision trees

Decision trees are a fundamental and extensively utilised technique in the field of machine learning and data analysis, renowned for their intuitive structure and interpretability [Myles et al., 2004, Navada et al., 2011, Rokach, 2016].

This hierarchical model recursively partitions the feature space into subsets, dividing the data based on the most informative features at each internal node. By repeatedly splitting the data according to the values of specific features, the model generates a tree-like structure that culminates in a prediction for the target variable. The tree's composition consists of nodes, where each internal node represents a decision based on the value of a particular feature, guiding the data down a specific branch of the tree. The leaf nodes at the end of these branches then denote the final output or class label, providing a transparent and easily interpretable prediction. This

structure allows decision trees to capture complex relationships in the data, making them a versatile and widely adopted tool for both classification and regression tasks in machine learning.

The construction of a decision tree involves a sophisticated process of selecting the most informative feature X_j and corresponding threshold t at each node to split the data. For a classification task, the criterion for selecting a feature can be based on information gain IG, which measures the reduction in entropy H after the split. For a set of data S, the information gain when splitting on feature X_j with threshold t can be defined as:

$$IG(S, X_j, t) = H(S) - \left(\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R) \right),$$

where S_L and S_R are the subsets of S resulting from the split. Entropy $H(S)$ is defined as:

$$H(S) = - \sum_{c \in \text{Classes}} p_c \log_2(p_c),$$

where p_c is the proportion of instances in S that belong to class c.

For regression tasks, the criterion for selecting the best split often involves minimizing the variance within the subsets. The variance reduction is given by:

$$\text{VarReduction}(S, X_j, t) = \text{Var}(S) - \left(\frac{|S_L|}{|S|} \text{Var}(S_L) + \frac{|S_R|}{|S|} \text{Var}(S_R) \right),$$

where $\text{Var}(S)$ is the variance of the target values in the set S. These carefully chosen criteria aim to create the most homogeneous possible subsets of the data, thereby enhancing the overall accuracy and predictive power of the decision tree model.

The recursive splitting of the decision tree continues until a stopping criterion is met, which serves to limit the model's complexity and prevent overfitting. This stopping criterion can take several forms, such as reaching a predefined maximum permissible tree depth, ensuring a minimum number of samples in each node, or achieving pure leaf nodes that contain only instances belonging to a single class. By imposing these constraints, the decision tree algorithm balances the trade-off between model complexity and generalization performance, ensuring that the final tree structure captures the underlying patterns in the data without becoming excessively intricate and prone to overfitting.

A key advantage of decision trees is their unparalleled interpretability. Unlike many other opaque machine learning models, decision trees offer remarkable transparency and ease of understanding. The hierarchical structure of a decision tree, with its internal nodes representing decisions based on feature values and the leaf nodes denoting the final output, allows for a clear, step-by-step explanation of how the model arrives at its predictions. Each path from the root to a leaf can be readily interpreted as a sequence of logical rules, making the decision-making process highly intuitive and accessible, even to those without extensive expertise in complex mathematical modelling

techniques. This remarkable interpretability is a crucial asset in applications where transparency and accountability are of paramount importance, such as in medical diagnosis, credit risk assessment, or regulatory compliance, where stakeholders require a clear understanding of the model's reasoning. The interpretability of decision trees enables users to gain valuable insights into the underlying patterns and relationships within the data, empowering them to make informed decisions with confidence. Despite their advantages, decision trees also have limitations that must be carefully considered. They are particularly prone to overfitting, a common issue that arises when the decision tree grows excessively deep. In such cases, the model may start to capture noise and idiosyncratic patterns in the training data, rather than learning the underlying relationships that are generalizable to new, unseen data. This can lead to poor performance on independent test sets, as the model fails to generalize effectively. However, there are several techniques that can be employed to mitigate the overfitting problem and enhance the decision tree's generalization capabilities. Pruning, which involves selectively removing branches of the tree to simplify the model, can be an effective strategy to prevent the tree from becoming overly complex and fitting to noise. Setting a maximum depth, or the maximum number of levels in the decision tree, is another approach that can help control the model's complexity and avoid overfitting.

Furthermore, ensemble methods, such as random forests and gradient boosting, have proven to be powerful tools for improving the performance of decision trees. These techniques combine multiple decision trees, each trained on a different subset of the data or with a different set of parameters, to create a more robust and accurate predictive model. By leveraging the collective strength of multiple decision trees, ensemble methods can overcome the limitations of individual trees and provide superior generalization performance [Dietterich, 2000, Natekin and Knoll, 2013, Rokach, 2016].

Ensemble methods

Ensemble methods represent a powerful class of machine learning techniques that aggregate the predictions of multiple models to generate a single, often more accurate, prediction. The fundamental advantage behind ensemble methods is that by aggregating the predictions of several models, the ensemble can diminish the variance, bias, or enhance the generalisation capability of the final model, compared to individual models. This is achieved by harnessing the diversity of the individual models, where their unique strengths and weaknesses are leveraged to produce a more robust and reliable prediction. Ensemble techniques, particularly those based on decision trees such as bagging and boosting, have emerged as some of the most powerful and versatile tools in the machine learning toolbox González et al. [2020].

One of the most ubiquitous ensemble techniques is Bagging, short for *Bootstrap Aggregating*. Bagging aims to reduce the variance of a model by training multiple instances of a model on different subsets of the data and then averaging their predictions. Each model in the ensemble is trained

on a bootstrap sample, which is a random sample of the original dataset drawn with replacement. This implies that some data points may appear multiple times in a bootstrap sample while others may not appear at all. This process introduces diversity among the individual models, as they are trained on slightly different subsets of the data. Once each model is trained, their predictions are aggregated, typically by averaging in the case of regression or by majority voting in the case of classification.

A prominent example of bagging is the random forest algorithm, which extends the concept of bagging to decision trees. Considering a dataset $D = \{(X_i, y_i)\}_{i=1}^{n}$, a random forest constructs a number B of decision trees, each trained on a bootstrap sample D_b of the original dataset. For a given input X, the prediction $\hat{y}$ of the random forest is obtained by averaging the predictions $\hat{y}_b$ of the individual trees (in regression) or by majority voting (in classification):

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b.$$

During the training of each tree, at each node, a random subset of m features is selected from the total p features, and the best feature from this subset is used to split the node. This random selection of features further reduces the correlation between trees, leading to a more robust model. random forests are highly effective as they mitigate two key issues associated with decision trees: overfitting and instability. By aggregating the predictions of multiple decision trees, the random forest approach diminishes the likelihood of overfitting, as any individual trees that may excessively fit the training data are counterbalanced by others. Additionally, the randomisation introduced through both bootstrap sampling and random feature selection at each split decreases the model's variance, rendering random forests less susceptible to noise within the data. However, aggregation is not the sole approach to model ensembling.

Boosting is a powerful ensemble technique that focuses on converting a set of weak learners into a strong learner. Unlike bagging, where models are trained independently, and their predictions are averaged, boosting trains models sequentially. Each model in the sequence attempts to correct the errors of its predecessor, focusing on the misclassified or poorly predicted cases from the previous model. This iterative process results in a model that gradually improves its performance as it learns to better handle the more challenging data instances that earlier models struggled with. By concentrating on the difficult cases, the boosting algorithm is able to incrementally enhance the overall predictive capability of the ensemble.

Mathematically, boosting can be described as a process that minimizes a specific loss function by adding models to the ensemble in a greedy fashion. At each iteration t, a new model $h_t(X)$ is added to the ensemble, and the model's contribution is weighted by a factor α_t, which is determined based on the model's performance. The overall prediction of the boosted model after T

iterations is given by:

$$\hat{y} = \sum_{t=1}^{T} \alpha_t h_t(X).$$

One of the most widely used boosting algorithms is adaptive boosting (AdaBoost) [Anghel et al., 2018]. AdaBoost iteratively assigns higher weights to training examples that were misclassified by the previous weak learner, forcing the next model to focus more on those problematic instances. Gradient boosting is another popular boosting technique, proposed by Friedman. This method generalizes the AdaBoost approach by allowing the use of arbitrary loss functions rather than being limited to the exponential loss used by AdaBoost.

Extreme gradient boosting (XGBoost) [Mitchell and Frank, 2017] is a highly efficient and scalable implementation of the gradient boosting algorithm. Due to its impressive performance across a wide range of problems, it has become a go-to choice for many machine learning practitioners. XGBoost is an advanced implementation of the gradient boosting algorithm that employs various optimisations to enhance its performance. It utilises a technique called shrinkage, or learning rate, which scales down the contribution of each individual tree, allowing for the addition of more trees to the ensemble, ultimately improving the model's accuracy.

Additionally, XGBoost incorporates column subsampling, similar to the random forests algorithm, where a random subset of features is considered at each split. This reduces overfitting and enhances the model's ability to generalise to unseen data. Furthermore, XGBoost is equipped to handle missing data by automatically learning the optimal approach to address such gaps during the training process. This is a significant advantage, as many real-world datasets often contain missing values that can pose challenges for other algorithms. Moreover, the XGBoost algorithm is designed to maximise computational efficiency through parallel and distributed computing, rendering it scalable and capable of handling large datasets. These combined techniques make XGBoost both rapid and robust, enabling it to frequently outperform other algorithms in practical applications across a wide range of tasks, such as classification, regression, and ranking.

In XGBoost, the objective function to be minimized is designed to balance the accuracy of the model with its complexity, thereby preventing overfitting. The objective function $Obj(\theta)$ at iteration t consists of two primary components: the loss function L, which measures how well the model's predictions fit the observed data, and a regularization term Ω, which penalizes model complexity. This objective function can be expressed as:

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t).$$

Here, y_i denotes the true label for instance i, $\hat{y}_i = \sum_{t=1}^{T} f_t(X_i)$ represents the predicted value, where the prediction is the sum of the outputs from all trees, and $f_t(X_i)$ is the output of the new model f_t being added at iteration t.

The regularization term $\Omega(f_t)$ is designed to control the complexity of the model by penalizing the complexity of the trees in the ensemble. For tree-based models in XGBoost, the regularization term is defined as:

$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_j^2$$

where T_t is the number of leaves in the t-th tree, γ is a regularization parameter that penalizes the number of leaves (and thus the complexity of the tree), and λ is a regularization parameter that penalizes the $L2$ norm of the leaf weights w_j.

At each iteration t, XGBoost adds a new tree $f_t(X)$ to minimize the objective function. The prediction after the t-th iteration is updated as follows:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(X_i)$$

The function $f_t(X)$ is chosen to minimize the following approximation of the objective function:

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(X_i) + \frac{1}{2} h_i f_t(X_i)^2 \right] + \Omega(f_t)$$

where $g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}$ is the first-order derivative of the loss function with respect to the prediction, and $h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$ is the second-order derivative of the loss function with respect to the prediction.

Each tree $f_t(X)$ in XGBoost is constructed by splitting the data based on feature values. The structure of a tree can be defined as:

$$f_t(X) = w_{q(X)}$$

where q is a function that maps each input X to a specific leaf in the tree, and $w_{q(X)}$ is the weight associated with the leaf to which X is assigned. The goal during training is to find the structure of q (i.e., the splits) and the corresponding weights w_j that minimize the objective function.

To determine the best split at each node during tree construction, XGBoost uses a measure called the "gain," which is the improvement in the objective function from making a split:

$$\text{Gain} = \frac{1}{2} \left[\frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \frac{G_L^2}{H_L + \lambda} - \frac{G_R^2}{H_R + \lambda} \right] - \gamma$$

where G_L and G_R are the sums of the first-order gradients for the left and right child nodes, respectively, H_L and H_R are the sums of the second-order gradients for the left and right child nodes, respectively, λ is the regularization parameter for the leaf weights, and γ is the regularization parameter for the number of leaves.

After training, the final prediction for a new instance X_i is obtained by summing the predictions of all the trees:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(X_i)$$

This ensemble of trees, each contributing to the final prediction, allows XGBoost to achieve high accuracy while maintaining control over model complexity. In summary, XGBoost builds an ensemble of decision trees by iteratively adding trees that correct the errors of the combined ensemble. It uses a combination of first- and second-order gradient information to guide the construction of each tree and incorporates regularization to prevent overfitting. The result is a powerful and scalable model capable of handling a wide range of machine-learning tasks.

2.1.3 Unsupervised learning

Unsupervised learning is a core discipline within machine learning that concentrates on unveiling concealed patterns, structures, or connections in data without relying on explicit labels or pre-determined outcomes. In contrast to supervised learning, where models are trained on labelled data to forecast a specific target variable, unsupervised learning exclusively operates on the input data, endeavouring to infer the inherent structure or distribution. The central aim of unsupervised learning is to uncover the inherent characteristics of the data and to identify pertinent clusters, connections, or dimensionality reductions that can provide insights into the data's underlying structure. This approach is especially valuable when working with large and complex datasets where the relationships between variables are not readily apparent [Steinbach et al., 2004].

One of the most prevalent tasks in unsupervised learning is clustering, where the goal is to group comparable data points together based on specific attributes. This technique aims to partition the data into meaningful subgroups, allowing researchers to uncover inherent structures and relationships that may not be immediately apparent.

K-means clustering

K-means clustering is a popular unsupervised machine learning algorithm designed to partition a dataset into K distinct clusters. The algorithm iteratively assigns data points to the nearest centroid and updates the centroids to the mean of the points assigned to each cluster. This process continues until the centroids converge, effectively grouping similar data points into well-defined clusters. Its simplicity and computational efficiency make K-means a widely applied method for uncovering inherent structures within large, unlabelled datasets.

Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$, where each $x_i \in \mathbb{R}^d$ is a d-dimensional vector, the objective of K-means is to identify K centroids $\mu_1, \mu_2, \ldots, \mu_K$ that minimize the sum of squared distances between each data point and its nearest centroid. This can be expressed as:

$$J(\{\mu_k\}_{k=1}^{K}) = \sum_{i=1}^{n} \min_{k \in \{1,\ldots,K\}} \|x_i - \mu_k\|^2,$$

where $\|x_i - \mu_k\|^2$ is the squared Euclidean distance between x_i and μ_k.

The algorithm begins by randomly initializing K centroids from the dataset. Each data point is then assigned to the cluster corresponding to the nearest centroid. Afterward, the centroids are recalculated as the mean of the data points in each cluster. These two steps—cluster assignment and centroid update—are repeated until the centroids stabilize, indicating convergence.

K-means' combination of simplicity and efficiency has led to its widespread use across various domains, making it a fundamental tool for exploratory data analysis and pattern recognition.

K-nearest neighbors

k-nearest neighbors (KNN) algorithm is a simple yet powerful method used in both classification and regression tasks. It is a non-parametric, instance-based learning algorithm that relies on the distance between data points to make predictions. The central idea of KNN is to classify a given data point based on the majority class among its K nearest neighbours or, in the case of regression, to predict the output value as the average of the outputs of the K nearest neighbours. Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$, where each $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector, the KNN algorithm makes a prediction for a new instance x by first identifying the K nearest neighbours of x in the feature space. The proximity between data points is typically measured using a distance metric (e.g. Euclidean distance), defined as:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}.$$

For classification, the predicted class $\hat{y}$ for a new instance x is determined by the majority vote among the classes of its K nearest neighbours. Mathematically, this can be expressed as:

$$\hat{y} = \text{argmax}_c \sum_{i \in \mathcal{N}_K(x)} \mathbb{I}(y_i = c),$$

where $\mathcal{N}_K(x)$ denotes the set of indices corresponding to the K nearest neighbours of x, y_i is the class label of the i-th neighbour, and $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the argument is true and 0 otherwise. For regression, the predicted output $\hat{y}$ is computed as the mean of the outputs of the K nearest neighbors:

$$\hat{y} = \frac{1}{K} \sum_{i \in \mathcal{N}_K(x)} y_i.$$

The choice of K plays a crucial role in the performance of the KNN algorithm. A smaller value of K makes the model more sensitive to noise,

while a larger value may smooth out the decision boundary too much, leading to underfitting. KNN can also be adapted to handle missing values in the data, making it a versatile tool in situations where incomplete datasets are common. One approach to handling missing values in KNN is to modify the distance calculation to ignore the dimensions with missing values. In other words, the distance between two instances x_i and x_j is computed only over the dimensions where both instances have valid (non-missing) values. The modified distance metric can be expressed as:

$$d(x_i, x_j) = \sqrt{\frac{1}{|S|} \sum_{k \in S} (x_{ik} - x_{jk})^2},$$

where $S \subseteq \{1, \ldots, d\}$ is the set of dimensions where both x_i and x_j have non-missing values, and $|S|$ is the number of such dimensions. This approach ensures that the distance metric is still valid even when some data points have missing values, allowing the KNN algorithm to function effectively. Another approach is to use KNN to input missing values. In this method, for each instance with a missing value in a particular dimension, the missing value is imputed by taking the mean (for regression) or the mode (for classification) of that feature among the K nearest neighbours where the feature value is available. This iterative imputation method leverages the underlying structure of the data to provide more accurate estimates of missing values, thereby improving the overall quality of the dataset before performing further analysis or training. The ability to adapt the distance metric or impute missing values makes KNN a flexible and robust algorithm for real-world applications where data completeness cannot always be guaranteed.

Clustering is a key task in unsupervised learning, with several algorithms available to group data points based on their similarities. While K-means and KNN are widely used clustering algorithms due to their simplicity and efficiency, there are other algorithms that offer different advantages depending on the nature of the data.

Hierarchical clustering [Ward, 1963, Murtagh and Contreras, 2011], for example, builds a hierarchy of clusters, enabling the identification of nested groupings at different levels. This method is particularly useful when the data's structure is inherently hierarchical, allowing for a more nuanced exploration of relationships between clusters. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [Ester et al., 1996, Schubert et al., 2017], meanwhile, identifies clusters based on density, effectively handling datasets with varying densities and the presence of outliers. Unlike K-means, which requires specifying the number of clusters in advance, DBSCAN automatically determines the number of clusters and is robust to noise, making it well-suited for complex datasets where clusters are irregularly shaped or of differing densities. Another important aspect of unsupervised learning is dimensionality reduction [Ayesha et al., 2020], which is concerned with transforming high-dimensional data into a lower-dimensional representation while preserving the essential characteristics of the original data.

Techniques like *Principal Component Analysis* [Jolliffe, 2005, Jolliffe and

Cadima, 2016] and *t-Distributed Stochastic Neighbor Embedding* [van der Maaten and Hinton, 2008] are popular methods for accomplishing this goal. The challenges of unsupervised learning are significant, primarily because the absence of labelled data makes it difficult to evaluate the performance of models and to determine the correctness of the patterns discovered. However, unsupervised learning remains a powerful tool in exploratory data analysis and is increasingly used in conjunction with supervised learning to enhance the accuracy and interpretability of models.

Evaluation Metrics

Specific metrics are used to quantify the reliability of a machine learning model and establish its predictive abilities. Since regression and classification tasks are different, we will discuss the metrics related to each separately.

Regression Metrics: As previously mentioned, the output of a regression model is an estimate of the target quantity (in our case, the Transit Time of CMEs), starting from a vector representation of the event encoding its characteristics. The most natural way to measure how well a model maps inputs to outputs is to quantify the distance between the model's prediction and the actual value. It is possible to define more than one metric to characterise this information. A first example is the MAE defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{2.1}$$

where N is the number of test samples, y_i are the actual values and $\hat{y}_i$ are the models predictions. In simpler words, the MAE score is the average of the absolute error values. Another widely used metric is the MSE, which is the squared difference between actual and predicted value:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{2.2}$$

MSE penalises predictions far away from the actual value and is, therefore, more sensitive to outliers than MAE. In addition, the MSE returns a value that is a squared unit of output, which makes interpretation less straightforward. Typically, to obtain such a measure in the same units as the model output, the square root of the MSE is used:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{2.3}$$

One of the most widely used metrics, because it is independent of the context in which the model is applied, is the R squared (R^2). R^2 is a statistical quantity which measures to what extent the variance of one variable (model predictions) explains the variance of a second variable (actual values).

Formally:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}, \tag{2.4}$$

where $\bar{y}$ is the average value of y.
It is generally better to look at more than one metric to evaluate a regression model because each one returns a slightly different piece of information. What has just been said is even more true regarding classification problems.

Classification Metrics: There are several ways to determine how well a classification model does its job. By comparing the labels predicted by the model with the assigned labels (also known as 'true labels'), the outputs are defined as True Positive (TP) when the model correctly predicts an instance as positive, while they are defined as True Negative (TN) when the model correctly predicts an instance as negative.

Similarly, outputs are defined as FP when the model incorrectly predicts an instance as positive, while they are defined as False Negative (FN) when the model incorrectly predicts instance as non negative. Obviously, the objective of a classifier is to maximise the number of true positives and true negatives so as to have few cases of misclassification. The most common metric used to evaluate performance in a classification task is the accuracy, which represents the percentage of instances that the model correctly predicts out of the total number. It is formally expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.5}$$

However, when the classes are characterised by a strong imbalance, as in the case studied in this work, the accuracy value may not fully represent the model's ability to distinguish the two classes because of the disproportion in size. Basically, the classifier can achieve high accuracy values even if it correctly predicts all or most instances of the majority class but fails to predict the minority class. This is one reason why combining more than one metric to assess classification performance is often better for getting a complete view.

Another typical evaluation metric for binary classification problems is the Receiver Operating Characteristic (ROC) curve. The ROC curve is obtained by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for different Threshold values and is typically used to visualise the discriminative ability of a binary classifier. TPR and FPR are also known as *Sensitivity* and *Specificity*, defined as:

$$TPR = \frac{TP}{TP + FN}, \quad TNR = \frac{TN}{TN + FP}. \tag{2.6}$$

The Area Under the Curve (AUC) is the area of the TPR $\times$ FPR space below the curve which is a summary of the ROC curve. The higher the AUC value, the better is the classifier. Although ROC is also influenced by the imbalance of the problem, it is more reliable than accuracy in assessing the actual capabilities of the classifier. In addition, it is possible to define class-specific evaluation

metrics, Precision and Recall:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}. \tag{2.7}$$

In other words, Precision returns information about the number of instances predicted as positive, which are actually labelled as positive. Recall, on the other hand, returns information about the number of instances that are labelled as positive and are actually predicted as positive.

Typically, the information from Precision and Recall is condensed into the F1 score, which is a harmonic mean of the two.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{2.8}$$

Some metrics directly take class imbalance into account, one of which is Balanced accuracy, which is an average of Specificity and Sensitivity:

$$Balanced\ accuracy = \frac{Specificity + Sensitivity}{2}. \tag{2.9}$$

Balanced accuracy returns a score more representative of a classifier's capabilities when class imbalance is significant.

It is important to emphasise once again that the list of relevant metrics to evaluate a classifier is very long because the goodness of the classifier is highly dependent on the context of the problem.

Validation

Evaluating the capability of a machine learning model to accomplish a given task requires robust validation methods. The simplest and most commonly used technique involves splitting the available dataset into two distinct subsets: the training set and the test set. The training set is used to train the model, while the test set evaluates its performance. This approach ensures that the test set contains data points that the model has not encountered during training, making it an effective test-bed for assessing the model's generalisation ability. Typically, the train/test split is performed randomly to minimise the risk of introducing bias.

In the context of CAT-PUMA, a similar validation method is employed with a significant enhancement. After optimising the SVM model, it identifies the train/test split that yields the best performance score from 10^6 random splits. This method, referred to as Best Split Validation (BSV) (Best Split Validation), is designed to select the split that ensures the highest representativeness of the training data within the test set. By maximising the test set's representativeness, this method provides a promising way to evaluate the model's learning ability.

However, BSV has limitations, particularly in scenarios where data availability is constrained. The selected best split might introduce bias, leading to an overly optimistic evaluation of the model's performance. To address

this concern and provide a more robust evaluation, we complement BSV with k-fold Cross Validation (CV) (cross-validation), a widely used technique in machine learning [Refaeilzadeh et al., 2009, Yadav and Shukla, 2016].

The k-fold cross-validation method is a more conservative approach that mitigates the risks associated with single train/test splits. It involves the following steps:

1. A set of hyper-parameters is selected for the model.
2. The dataset is divided into k equal subsets (folds). The model is trained on $k-1$ folds, and the remaining fold serves as the validation set to evaluate performance. The performance score for this validation fold is recorded.
3. This process is repeated k times, with each fold being used once as the validation set and the remaining $k-1$ folds used for training. This ensures that the model is trained and evaluated on all subsets of the dataset.
4. Once all k iterations are completed, the average performance score across all validation sets is calculated and stored.
5. A new set of hyper-parameters is selected, and the process is repeated from step 1.

The k-fold cross-validation approach offers several advantages. By training the model on multiple combinations of the data and evaluating its performance on different validation sets, it reduces the variance associated with a single train/test split. This results in a more reliable and robust estimation of the model's performance. Moreover, it allows for a comprehensive evaluation of how the model generalises to unseen data, providing confidence in its applicability to real-world scenarios. In this thesis, the combination of BSV and k-fold CV ensures a balanced evaluation framework that accounts for both representativeness and robustness, addressing potential biases while leveraging the strengths of each method.

In addition to BSV and k-fold CV, it is worth noting that many other validation techniques exist, each tailored to specific needs and scenarios. For instance, holdout validation is often used when datasets are large and computational efficiency is a priority, as it involves splitting the data into distinct training, validation, and test sets [Hastie et al., 2005]. This method ensures that the test set remains untouched during the training and hyper-parameter tuning phases, providing a reliable estimate of the model's final performance on unseen data.

Nested cross-validation is another robust technique, particularly useful for scenarios involving hyper-parameter optimisation. It employs an inner loop for hyper-parameter tuning and an outer loop for performance evaluation [Cawley and Talbot, 2010]. This approach prevents data leakage and ensures that the performance metrics reflect the model's ability to generalise, even when extensive parameter tuning is required.

Furthermore, stratified k-fold cross-validation is particularly effective for imbalanced datasets, as it maintains the same class distribution in each fold as in the original dataset [Kohavi, 1995]. This ensures that the validation performance is not skewed due to class imbalance, making it an essential technique for classification problems involving rare events.

These additional methods highlight the importance of selecting a validation strategy that aligns with the specific requirements of the dataset and the problem at hand. By leveraging multiple validation approaches, this thesis ensures a comprehensive and rigorous evaluation of the proposed models, striking a balance between computational efficiency, reliability, and generalisation performance.

Hyperparameter Tuning

Hyperparameter tuning is a crucial step in optimising machine learning models to achieve the best performance for a given task. Hyperparameters are parameters that are not learned during the training process but are set prior to training and control aspects of the model's behaviour, such as learning rate, regularisation strength, or the number of layers in a neural network.

The tuning process involves exploring the space of hyperparameters to identify the combination that yields the best performance. Several methods exist for performing hyperparameter optimisation:

Grid Search: This method involves exhaustively searching through a predefined set of hyperparameter values. A grid is constructed where each dimension corresponds to a hyperparameter, and all possible combinations of hyperparameter values are evaluated. While this approach guarantees that the optimal combination within the predefined grid is identified, it can be computationally expensive, especially for high-dimensional parameter spaces.

Random Search: Unlike grid search, random search samples hyperparameter values randomly from specified distributions [Bergstra and Bengio, 2012]. This approach is more efficient for high-dimensional spaces, as it tends to explore a broader range of hyperparameter values and often finds good combinations faster than grid search.

Bayesian Optimisation: This advanced method uses a probabilistic model to approximate the objective function and guides the search process based on prior evaluations. One commonly used technique is the Tree-structured Parzen Estimator (Tree-structured Parzen Estimator (TPE)) [Bergstra et al., 2011]. TPE models the hyperparameter space and focuses the search in regions where high performance is more likely. This significantly reduces the number of evaluations required and is particularly effective for complex optimisation problems.

In this thesis, hyperparameter tuning is performed using TPE implemented with the Optuna[1] optimisation library [Akiba et al., 2019]. Optuna is an open-source framework that allows efficient exploration of the hyperparameter space by leveraging the history of previous trials to direct future searches. By

[1]Optuna documentation: https://optuna.readthedocs.io/en/stable/index.html

concentrating the search in regions with higher performance, this method accelerates the optimisation process and reduces computational overhead.

Hyperparameter tuning is essential for achieving robust and high-performing models. By systematically exploring and optimising hyperparameter values, we ensure that the models developed in this thesis are both efficient and effective, capable of addressing the specific challenges of the task at hand.

2.2 Bayesian inference of the parameters

In this section, we describe the theoretical background behind MCMC methods, which are widely used for sampling from complex probability distributions. MCMC methods rely on the principles of Markov chains, which are stochastic processes characterised by the property that the next state depends only on the current state. By carefully constructing a Markov chain that has the target distribution as its stationary distribution, MCMC enables efficient sampling even in high-dimensional or analytically intractable scenarios.

The MCMC approach is built upon the concept of Markov chains, which involve constructing a sequence of samples in the parameter space that progressively converges to a stationary distribution, corresponding to the target posterior probability distribution. To understand the foundation of MCMC methods, it is essential to first introduce the general properties and principles of Markov chains, which form the backbone of this sampling technique [Ivezić et al., 2019, Anzai, 2012].

2.2.1 Markov chains

The first step is to formalize the intuitive idea of these objects. A Markov chain is a sequence of random variables $X_1, X_2, ..., X_N$ such that the distribution of X_{i+1} depends only on the value of X_i, and not on the full history of prior values $X_1, ..., X_{i-1}$, i.e.,

$$\mathcal{P}(X_{i+1} = x_{i+1} | X_j = x_j, \{X_j, j = 0, ..., i-1\}) = \mathcal{P}(x_{i+1} | x_i). \tag{2.10}$$

In other words, the transition to a subsequent state depends only on the current state and not on the entire sequence of states leading to it. A stochastic process fulfilling this *Markov property* is termed a *Markov process*. The initial distribution, denoted by λ, describes the probability distribution of the initial state X_0:

$$\mathcal{P}(x_0) = \lambda(x_0), \tag{2.11}$$

which provides the marginal probability for the initial states of the process. The transition from one state x at step i to another state x' at step $i+1$ is characterized by a kernel function $\mathcal{T}_i(x, x')$, known as the transition kernel, which is the conditional probability $\mathcal{P}(X'|X)$ for transitioning from X to X'.

In a discrete Markov process, these transitions can be represented as a matrix $\mathcal{T}$.

The transition from state X_t to state X_{t+1} can be described by a transition matrix $\mathcal{T}$, where each element is defined as:

$$t_{ij} = \mathcal{P}(X_{t+1} = x_j | X_t = x_i), \tag{2.12}$$

with $t_{ij} \geq 0$ and $\sum_j t_{ij} = 1$ for all i. The probabilities for the initial states are given by $\lambda_i = \mathcal{P}(X_1 = x_i)$, with $\sum_i \lambda_i = 1$.

The Markov chain's transitions can also be visualized as a directed graph, where nodes represent states and edges between nodes represent transitions, labelled by the transition probabilities $\mathcal{P}(X_{t+1} = x_j | X_t = x_i)$. For a given Markov chain $\{X_n\}$ with a transition matrix $\mathcal{T}$, the n-step transition matrix, denoted by $\mathcal{T}^{(n)} = (t_{ij}^{(n)})$, describes the probability of transitioning from state x_i to state x_j after n steps. This n-step transition probability is defined as:

$$\mathcal{P}(X_{m+n} = x_j | X_m = x_i) = t_{ij}^{(n)}. \tag{2.13}$$

For a homogeneous (or stationary) Markov chain, where the transition probabilities are independent of time (i.e., $\mathcal{T} \equiv \mathcal{T}(x, x')$), the Markov chain is completely specified by the initial probabilities λ and the transition matrix $\mathcal{T}$. The evolution of the marginal probability distribution $p_i(x)$ for a given state x at step i can be expressed as:

$$p_i(x) = \sum_{x'} p_{i-1}(x') \mathcal{T}(x', x). \tag{2.14}$$

A probability distribution $\pi(x)$ is said to be invariant with respect to a Markov chain if it remains unchanged by the transition process:

$$\pi(x) = \sum_{x'} \pi(x') \mathcal{T}(x', x). \tag{2.15}$$

Once a Markov chain reaches this invariant distribution, it remains stationary. An important condition that ensures a distribution $\pi(x)$ is invariant is the detailed balance condition:

$$\pi(x) \mathcal{T}(x', x) = \pi(x') \mathcal{T}(x, x'). \tag{2.16}$$

If this condition is satisfied, the transition probabilities leave the distribution invariant, and the Markov chain is said to be reversible with respect to $\pi(x)$. In practice, MCMC methods leverage these properties to sample from a target distribution $\pi(x)$. The goal is to construct a Markov chain whose unique invariant distribution corresponds to the desired target distribution, often referred to as the *equilibrium distribution*.

For many applications, it is essential that as $i \to \infty$, the distribution $p_i(x)$ converges to the invariant distribution $\pi(x)$, regardless of the initial distribution $\lambda(x)$. In such cases, the Markov chain is called ergodic, meaning

that it eventually "forgets" its starting point and samples from the equilibrium distribution [2].

MCMC methods are particularly powerful due to this ergodic property. A well-constructed Markov chain, regardless of its initialization, will converge to a stationary distribution $\pi(x)$, allowing for efficient sampling from complex probability distributions. The following section is dedicated to a more in-depth understanding of the MCMC approach.

2.2.2 Monte carlo markov chains

The growing popularity of Bayesian inference in various scientific and economic fields has been driven by the increased feasibility of numerical simulations in recent years. Markov chain Monte Carlo methods, which originate from the field of statistical physics, have gained significant importance in statistics since the late 1980s, largely due to technological advancements that have enabled their widespread application [Brooks, 1998, Cappé and Robert, 2000, van Ravenzwaaij et al., 2016].

MCMC algorithms are a well-established class of Monte Carlo methods that generate a Markov chain with the desired invariant distribution. Deriving the stationary distribution of a Markov chain analytically can be challenging or, in many cases, infeasible. The result is that it is often necessary to investigate the stationary distribution numerically through simulation techniques. These powerful techniques have enabled the numerical mapping of posterior distributions, even in complex scenarios with high-dimensional parameter spaces and intricate structures featuring multiple peaks. MCMC methods aim to efficiently sample the posterior distribution, concentrating on regions with higher probability and excluding areas with low likelihood, thus providing a robust and flexible framework for probabilistic inference.

MCMC algorithms are designed with carefully constructed transition kernels, ensuring that the detailed balance condition is satisfied, allowing the chain to converge to the desired stationary distribution. To ensure that a Markov chain reaches a stationary distribution proportional to $\pi(X)$, the probability of arriving at a point $X_{t+1} \equiv \boldsymbol{\theta}_{t+1}$ must be proportional to $\pi(\boldsymbol{\theta}_{t+1})$, which can be expressed as:

$$\pi(\boldsymbol{\theta}_{t+1}) = \int \mathcal{T}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) \pi(\boldsymbol{\theta}_t) d\boldsymbol{\theta}_t,$$

where $\mathcal{T}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})$ is the transition kernel that governs the transition between states. This requirement is satisfied when the transition probability adheres to the detailed balance condition (Eq. 2.16), making the choice of the transition kernel $\mathcal{T}$ crucial. Each MCMC algorithm employs a different transition kernel suited to the specific problem at hand. The most widely known and utilized MCMC algorithm is the Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970]. This algorithm is a cornerstone of MCMC methods and

[2] This property is crucial in MCMC methods because it guarantees that the chain will converge to the target posterior distribution, $\mathcal{P}(\theta|\mathcal{D})$, in Bayesian inference.

plays a fundamental role in numerous applications where complex posterior distributions need to be sampled efficiently.

2.2.3 The Metropolis-Hastings algorithm

The Metropolis-Hastings (M-H) algorithm, initially developed by Metropolis et al. [1953] and later generalized by Hastings [1970], is one of the most common MCMC algorithms. Its primary purpose is to generate a sequence of states that follow a target posterior distribution, thus facilitating efficient sampling in complex parameter spaces.

Let $X_t = \boldsymbol{\theta}_t \equiv (\theta_1^{(t)}, \ldots, \theta_m^{(t)})$ represent the state of the Markov chain at time t in an m-dimensional parameter space. The transition from the state X_t to X_{t+1} is governed by a transition probability, denoted by $\mathcal{T}(X_t, X_{t+1})$. This transition kernel is chosen in such a way that the Markov chain converges asymptotically to the desired stationary distribution $\pi(\boldsymbol{\theta})$.

The evolution of the Markov chain proceeds iteratively. At each step, a new state X_{t+1} is proposed based on a *Proposal Density*, $\mathcal{Q}(X_{t+1}|X_t)$, which depends only on the current state X_t. The proposed state is then either accepted or rejected based on a probability given by the acceptance ratio:

$$\alpha(X_t, X_{t+1}) = \min\left(1, \frac{\mathcal{Q}(X_t|X_{t+1})\pi(X_{t+1})}{\mathcal{Q}(X_{t+1}|X_t)\pi(X_t)}\right). \tag{2.17}$$

In this framework, the transition probability is expressed as:

$$\mathcal{T}(X_t, X_{t+1}) = \alpha(X_t, X_{t+1})\mathcal{Q}(X_{t+1}|X_t),$$

where the acceptance probability α acts as a correction factor to account for any discrepancy between the proposal density and the target distribution. This acceptance rule depends only on the ratio of posterior probabilities, allowing the algorithm to function even when the normalization constant of the target distribution is unknown.

The original version of the Metropolis algorithm employs a symmetric proposal distribution, i.e., $\mathcal{Q}(X_{t+1}|X_t) = \mathcal{Q}(X_t|X_{t+1})$, simplifying the acceptance ratio to the posterior ratio:

$$\alpha(X_t, X_{t+1}) = \min\left(1, \frac{\pi(X_{t+1})}{\pi(X_t)}\right).$$

In this case, the proposed state X_{t+1} is always accepted if $\pi(X_{t+1}) > \pi(X_t)$. If not, the proposed state is accepted with probability $\pi(X_{t+1})/\pi(X_t)$. This version of the algorithm satisfies the detailed balance condition, which ensures that the stationary distribution of the chain is indeed the target distribution $\pi(X)$. Mathematically, this is expressed as:

$$\pi(X_{t+1})\mathcal{T}(X_{t+1}, X_t) = \pi(X_t)\mathcal{T}(X_t, X_{t+1}),$$

guaranteeing that the equilibrium distribution of the chain is $\pi(X) \equiv \pi(\boldsymbol{\theta})$.

The performance of the Metropolis-Hastings algorithm heavily depends on the choice of the proposal density $\mathcal{Q}(X_{t+1}|X_t)$. While any distribution that moves through the parameter space can be used, the computational efficiency of the algorithm varies significantly depending on this choice. A common choice is to use a Gaussian distribution centred at the current state X_t, though more sophisticated proposals may be employed in practice to improve convergence rates.

2.2.4 Revised Metropolis-Hastings approach

The basic idea behind MCMC is an iterative procedure that creates a chain of values in the parameter space. Each iteration updates the parameter value according to a specific rule of acceptance, ensuring that the final distribution of the chain follows the target probability distribution. This allows for the exploration of complex, high-dimensional probability distributions that are often intractable to analyze using traditional analytical methods.

Probabilistic inference through Markov chains involves constructing sequences of points in the parameter space, where the density of these points is proportional to the target a-posteriori probability distribution. Notably, there exist Markov chains that converge to a single, stationary probability distribution, which can then be used to estimate the relevant statistical quantities, such as means, variances, and credible intervals, providing a powerful tool for a wide range of applications in science, engineering, and economics.

The strength of these Bayesian methods lies in their ability to explore the parameter space in search of the zone that best represents the observations in terms of likelihood. The algorithm employed in this work is a revisited version of the M-H algorithm, which can be summarised in the following steps:

- A parameter set $\boldsymbol{\theta}_t = (\gamma_t, w_t)$ is sampled from the parameter space, using a proposal distribution centred around the values sampled at the previous step, $\boldsymbol{\theta}_{t-1} = (\gamma_{t-1}, w_{t-1})$. The initial parameter set is sampled from the prior.

- The proposed set of parameters $\boldsymbol{\theta}_t$ is used to solve the DBM equations (1.6, 1.7) and obtain estimates of the arrival time and velocity $(\hat{\text{ToA}}, \hat{\text{VoA}})$ of the CME events in the dataset.

- The acceptance probability α is calculated using the Metropolis-Hastings ratio:

$$\alpha = min\left(1, \frac{\pi(\mathcal{D}|\boldsymbol{\theta}_t)}{\pi(\mathcal{D}|\boldsymbol{\theta}_{t-1})}\right), \tag{2.18}$$

  where

$$\pi(\mathcal{D}|\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) \times prior(\boldsymbol{\theta}). \tag{2.19}$$

Here, $\pi(\mathcal{D}|\boldsymbol{\theta})$ depends on the likelihood function ($\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$) and the prior distribution of the parameters. The substantial difference between the employed

MCMC algorithm and the standard M-H lies in the Likelihood function used to determine how well the proposed parameters are in accordance with the collected data. In particular, the probability of acceptance, and thus essentially the Likelihood function, represents the heart of this technique and is worth explaining in more detail. The likelihood function assesses the agreement between the result obtained with the proposed parameters and the observed data, which is the key component of the technique.

The *a-priori* distribution incorporates the previous knowledge about the parameters. By exploring the parameter space guided by the likelihood function, the MCMC algorithm efficiently explores the parameter space and constructs a so-called posterior distribution. The main idea here is to find a distribution for the DBM parameters that are valid to represent the observations of all CMEs contained in the dataset (or belonging to an ensemble with specific characteristics, such as accelerated or decelerated CMEs). To take this into account, the likelihood function for a set of parameters $\boldsymbol{\theta}$ given an ensemble $\mathcal{G}$ of CMEs is defined as the product of the individual likelihoods associated with each CME event. Each individual likelihood is proportional to a bivariate normal distribution centred on the observed (ToA, VoA) values. Hence, for a sampled set (γ, w) and an Ensemble $\mathcal{G}$ of CMEs (e.g. slow solar wind speed CMEs or fast solar wind speed CMEs), we write the likelihood function as:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{L}_{\mathcal{G}}(\gamma, w) = \prod_{cme \in \mathcal{G}} \mathcal{N}\left(\begin{bmatrix}\mathrm{ToA}_{cme}\\ \mathrm{VoA}_{cme}\end{bmatrix}, \Sigma_{cme}\right)\left(\begin{bmatrix}\hat{\mathrm{ToA}}_{cme}\\ \hat{\mathrm{VoA}}_{cme}\end{bmatrix}\right), \quad (2.20)$$

$$\Sigma_{cme} = \begin{bmatrix}\mathrm{Var}[\mathrm{ToA}_{cme}], & \mathrm{Cov}[\mathrm{ToA}_{cme}, \mathrm{VoA}_{cme}]\\ \mathrm{Cov}[\mathrm{ToA}_{cme}, & \mathrm{VoA}_{cme}], \mathrm{Var}[\mathrm{VoA}_{cme}]\end{bmatrix}, \quad (2.21)$$

where $\mathcal{N}$ represents a bivariate normal distribution with mean values (ToA_{cme}, VoA_{cme}) and covariance matrix Σ_{cme}, evaluated in the estimates ($\hat{\mathrm{ToA}}_{cme}, \hat{\mathrm{VoA}}_{cme}$) obtained solving the DBM equations with the proposed parameter set $\boldsymbol{\theta}_t$.

The covariance matrix Σ_{cme} in equation 2.21 captures the uncertainties in the observed values, allowing for deviations up to 10% of the observed values ($\mathrm{Var}[\mathrm{ToA}_{cme}] = 0.10 \times (\mathrm{ToA}_{cme})^2$, $\mathrm{Var}[\mathrm{VoA}_{cme}] = 0.10 \times (\mathrm{VoA}_{cme})^2$). They should ideally be equal to the estimated error measure, but to allow an easier MCMC method convergence, we allow errors up to 10% of the observed values. We tested the 10% threshold and found it to be a robust compromise between convergence and acceptance rate.

The anti-diagonal coefficient of Σ_{cme} accounts for the covariance between ToA and VoA that, in this case, is taken as the empirical correlation obtained from our data set and then scaled by the square root of the diagonal coefficient ($\mathrm{Cov}[\mathrm{ToA}_{cme}, \mathrm{VoA}_{cme}] = \mathrm{Corr}[\mathrm{ToA}_{cme}, \mathrm{VoA}_{cme}] \times \sqrt{\mathrm{Var}[\mathrm{ToA}_{cme}]}\sqrt{\mathrm{Var}[\mathrm{VoA}_{cme}]}$). To simplify computations, we utilize the log-likelihood to convert products of exponentials into sums of their respective arguments. The MCMC method allows for the incorporation of prior information on the parameters through the prior distribution term $\pi(\boldsymbol{\theta})$ in the acceptance probability calculation. In this study, we utilized uniform

(hence non-informative) prior distributions with boundaries extending well beyond physically plausible values for w and γ ($w \in [0, 1000][km/s]$ and $\gamma \in [0, 10^{-7}][km^{-1}]$). Using non-informative priors ensures that the posterior distributions are not influenced by specific prior assumptions, enabling an objective comparison with previous results.

The MCMC algorithm in this work includes the uncertainty in the travelled distance by incorporating it as a free parameter with a uniform prior distribution ($R \in [0.97, 1.20][AU]$). The algorithm is designed to accept candidate parameters only if they can solve the DBM equations for all CMEs in the ensemble.

This approach, referred to as the *ensemble approach*, provides parameter distributions representative of an ensemble of CMEs, allowing for modelling the interplanetary propagation of all CMEs belonging to that Ensemble.

Additionally, we developed an alternative version of the algorithm, referred to as the *individual approach*, that returns parameter distributions for each CME in the dataset independently. Before describing the results, it is important to highlight the methods used to assess the convergence of the algorithm and ensure the reliability of the obtained posterior distributions.

2.2.5 Convergence diagnostic

Valid inferences from MCMC samples rely on the assumption that the samples accurately represent the true posterior distribution. While theoretical guarantees ensure convergence to the target distribution as iterations approach infinity, determining the minimum number of iterations required for a sufficiently accurate approximation remains a practical challenge. This threshold varies depending on the problem, necessitating independent convergence assessments for each MCMC application.

Convergence diagnostics are essential tools for evaluating whether MCMC chains have reached the target distribution. A widely used diagnostic is the Gelman-Rubin method [Gelman and Rubin, 1992], which compares the variance between chains (*between-chain variance*) to the variance within each chain (*within-chain variance*). The diagnostic outputs the *R statistic*, also known as the Gelman-Rubin diagnostic, which assesses the convergence of chains by examining whether their variances are consistent. For K independent chains of length N, denoted as:

$$X_K = \left\{ \left(\theta_{1,k}^{(i)}, \ldots, \theta_{m,k}^{(i)} \right) \right\}_{i=1}^{N}, \quad k = 1, \ldots, K,$$

the chain-specific mean $\hat{\theta}_k$, overall mean $\hat{\theta}$, and variance $\hat{\sigma}_k^2$ are calculated as:

$$\hat{\theta}_k = \frac{1}{N} \sum_{i=1}^{N} \theta_k^{(i)}, \quad \hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k, \quad \hat{\sigma}_k^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\theta_k^{(i)} - \hat{\theta}_k \right)^2.$$

The between-chain variance B and within-chain variance W are:

$$B = \frac{N}{K-1}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta})^2, \quad W = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_k^2.$$

The pooled variance $\hat{V}$ is then computed as:

$$\hat{V} = \frac{N-1}{N}W + \frac{K+1}{KN}B,$$

and the potential scale reduction factor (Potential Scale Reduction Factor (PSRF)) is defined as:

$$PSRF = \sqrt{\frac{\hat{V}}{W}}.$$

A PSRF close to 1 indicates convergence, while significantly larger values suggest additional iterations are required.

2.2.6 Autocorrelation time

MCMC algorithms, including the Metropolis-Hastings method, generate chains of samples that converge to the stationary posterior distribution. However, due to the random walk nature of these methods, even after reaching the target region of parameter space, successive samples may exhibit autocorrelation, which reduces their independence.

The auto-correlation function (ACF) quantifies the dependence between samples in a Markov chain. It is defined as:

$$ACF(\tau) = \rho_{X,X_\tau} = \frac{\mathrm{cov}(X, X_\tau)}{\sigma\sigma_\tau} = \frac{\mathbb{E}[(X-\mu)(X_\tau - \mu_\tau)]}{\sigma\sigma_\tau},$$

where X_τ represents the delayed chain, and μ, σ are the mean and standard deviation of X and X_τ. The autocorrelation time, τ_{ac}, is the lag after which the autocorrelation function effectively approaches zero. Practically, τ_{ac} is the lag beyond which:

$$ACF(\tau) < \epsilon, \quad \text{for} \quad \tau > \tau_{ac},$$

where ϵ is a predefined threshold.

Once τ_{ac} is determined, a post-processing step called *thinning* can be applied to reduce sample correlation. Thinning involves retaining only one sample every τ_{ac} steps, ensuring the retained samples are approximately independent. Although thinning reduces the effective sample size, it improves the quality of the samples, making them more suitable for statistical inference and posterior summarization.

2.3 Mathematical morphology

MM is a non-linear image processing technique grounded in set theory, topology, and lattice algebra, designed for analyzing geometrical structures in images. MM was developed initially by Georges Matheron and Jean Serra in the 1960s to address challenges in binary image processing related to mining and materials sciences [Matheron, 1967, Serra, 1969], and has since been extended to grayscale and multivariate images. This method probes and transforms images by interacting them with a small predefined shape called a structuring element (SE). Through this probing, MM extracts critical information about spatial and geometrical properties such as size, shape, and topology.

In recent years, MM has been successfully applied to various fields, including medical imaging [Soille, 1999], materials science [Serra, 1983], and more recently astrophysics, particularly solar physics. In this domain, MM plays a key role in detecting, segmenting, and analyzing dynamic solar features such as filaments, sunspots, and facular regions [Shih and Kowalski, 2003, Barata et al., 2018, Bourgeois et al., 2025]. At the heart of MM are several key operations: erosion, dilation, opening, and closing, which are mathematically defined for binary and grayscale images. Let $A \subseteq E$ be a binary image, where A is the set of foreground pixels and E is the image domain. Let $B \subseteq E$ be the SE, a small, typically simple shape (e.g., a disk or line segment).

The erosion of set A by structuring element B, denoted by $A \ominus B$, is defined as:

$$A \ominus B = \{z \in E : B_z \subseteq A\},$$

where B_z is the translation of B by z, i.e., $B_z = \{b + z : b \in B\}$. Erosion shrinks the object by removing boundary pixels, effectively reducing the size of foreground regions. For grayscale images, erosion is extended as:

$$(A \ominus B)(x) = \min_{b \in B}\{A(x + b)\},$$

where $A(x)$ represents the intensity value at pixel x [Soille, 1999]. Erosion is useful for eliminating small or narrow structures in solar images, such as fine strands or weakly illuminated regions.

The dilation of set A by structuring element B, denoted $A \oplus B$, is defined as:

$$A \oplus B = \{z \in E : (B^s)_z \cap A \neq \emptyset\},$$

where B^s is the symmetric reflection of B, i.e., $B^s = \{-b : b \in B\}$. Dilation expands the object by adding pixels to the boundary. In grayscale, dilation is defined as:

$$(A \oplus B)(x) = \max_{b \in B}\{A(x - b)\}.$$

Dilation fills small gaps and connects nearby structures, making it valuable for consolidating fragmented solar features like active regions or sunspot boundaries [Soille, 1999].

The opening operation, denoted $A \circ B$, is defined as erosion followed by dilation:

$$A \circ B = (A \ominus B) \oplus B.$$

Opening smooths object boundaries, removes small noise elements, and separates objects that are close together. For grayscale images, opening removes bright small-scale structures [Soille, 1999]. The closing operation, denoted $A \bullet B$, is dilation followed by erosion:

$$A \bullet B = (A \oplus B) \ominus B.$$

Closing fills small holes within an object while preserving its general shape. In grayscale, the closing can remove small dark regions or local depressions, making it useful for enhancing solar features like coronal holes and sunspots [Serra, 1983].

For grayscale images where pixel intensities are real numbers $I(x) \in \mathbb{R}$, MM operations generalize by replacing set operations with point-wise infimum (min) and supremum (max). The dilation and erosion for grayscale images are defined by:

$$(I \oplus B)(x) = \sup_{b \in B} \{I(x - b) + B(b)\},$$

and

$$(I \ominus B)(x) = \inf_{b \in B} \{I(x + b) - B(b)\}.$$

This allows MM to process continuous-valued images, such as those obtained from solar observations, making it highly applicable for analyzing images with complex intensity variations, such as those from the SDO[Lemen et al., 2012].

Beyond basic operations, MM offers a variety of advanced operators for more specialized image analysis, such as morphological gradient, top-hat transform and skeletonization: The morphological gradient of a set A is defined as the difference between the dilation and erosion of the set:

$$\text{Grad}(A) = (A \oplus B) - (A \ominus B).$$

This operation highlights edges, which is particularly useful in detecting the boundaries of solar structures like filaments, sunspots, and coronal loops [Gonzalez, 2009].

The top-hat transform enhances small objects by subtracting the result of an opening or closing from the original image:

$$\text{Top-hat}(A) = A - (A \circ B).$$

This transform is effective in detecting small bright features such as coronal bright points or spicules [Soille, 1999].

Skeletonization reduces an object to its minimal representation while retaining topological features. The skeleton $\text{Skel}(A)$ of an object A can be

computed iteratively by applying erosion followed by subtraction:

$$\text{Skel}(A) = \bigcup_{k=0}^{\infty} (A \ominus B_k) - ((A \ominus B_k) \oplus B).$$

Skeletonization is crucial for analyzing complex solar structures like coronal loops or prominence threads [Vincent and Soille, 1991]. MM is extensively used in solar physics to detect, segment, and track solar features in noisy datasets. For example:

- Filament detection and tracking: MM is used to detect solar filaments by combining region-growing algorithms with morphological opening and closing operations to segment filamentary structures from surrounding corona [Shih and Kowalski, 2003, Bourgeois et al., 2025].
- Sunspot segmentation: MM aids in distinguishing between the umbra and penumbra regions of sunspots, enabling precise measurement of sunspot areas and the tracking of their evolution [Shih and Kowalski, 2003].
- Coronal jet identification: MM, in combination with semi-automated algorithms, helps detect small-scale coronal jets by enhancing weak features in solar images, which can be obscured by noise [Liu et al., 2023].

Recently, MM has been integrated with machine learning techniques to enhance feature detection and classification in solar data. By combining MM's ability to extract geometric information with machine learning's pattern recognition capabilities, researchers can automate the identification and analysis of solar phenomena at unprecedented scales. For example, DL algorithms that incorporate MM for preprocessing have been used to track filaments and detect jets with improved accuracy [Davidson and Ritter, 1990, Derivaux et al., 2007, Nogueira et al., 2019, Franchi et al., 2020, Mondal et al., 2020, Roy et al., 2021].

This chapter has outlined the core methodologies applied throughout the thesis, ranging from probabilistic physics-based modelling to machine learning techniques and image processing tools. Table 2.1 summarises the objectives, advantages, and limitations of the methods employed.

TABLE 2.1: Summary of methods used in this thesis: their purpose, key advantages, and limitations.

Method	Purpose	Advantages	Limitations
Machine Learning (SVM, RF, XGBoost)	Supervised learning for CME transit prediction, Earth-impact classification, and jet identification	Adaptable to mixed feature types; captures nonlinear relationships; supports automation across tasks	Requires careful tuning and preprocessing; may be sensitive to class imbalance or feature sparsity
SHAP	Post-hoc interpretation of machine learning models	Provides model-agnostic feature importance; increases transparency in complex ML systems	Adds computational overhead; explanations depend on reliable input-output mapping
P-DBM	CME transit time forecasting with uncertainty	Simplified Physics model; computationally efficient; allows probabilistic forecasting from minimal input parameters	Relies on steady solar wind assumptions; limited capacity to capture complex propagation dynamics
MCMC	Bayesian estimation of P-DBM parameters	Enables principled uncertainty quantification; incorporates prior knowledge and observational constraints	Sensitive to prior assumptions and sampling configuration; computationally demanding
Mathematical Morphology	Segmentation and shape-based feature extraction from solar limb images	Well-suited for structural filtering; enhances detection of geometrically coherent features	Sensitive to structuring parameters; limited robustness under high image noise or ambiguity

Chapter 3

Data

This chapter presents the datasets employed in this thesis, which underpin the analyses and models developed across its three main projects. Each dataset is meticulously curated to ensure reliability and relevance for the corresponding research objectives.

3.1 Earth-impacting CMEs dataset

The dataset of Earth-impacting CMEs was compiled following the approach outlined in Liu et al. [2018], leveraging multiple established CME catalogues:

- The Richardson and Cane List [Richardson and Cane, 2010],
- The full halo CMEs list from the University of Science and Technology of China [Shen et al., 2013],
- The George Mason University CME/ICME List [Hess and Zhang, 2017],
- The CME Scoreboard (NASA).

Essentially, the data mining process comprises two stages. First, we identify all the observed geoeffective CME events from 1996 to 2022, and then we associate each event with the features that will form the input space for the ML models. Ambiguous events and duplicates were excluded, resulting in a clean dataset of 324 CME events.

Features

The dataset comprises 17 features describing CMEs and solar wind states. CMEs features include velocity, mass, angular width, and the Measurement Position Angle (MPA), derived from the SOHO LASCO catalogue[1]. Solar wind features, such as plasma density, temperature, and magnetic field components in Geocentric Solar Ecliptic System (GSE) coordinates, were extracted from OMNIWeb Plus[2], averaged over a 6-hour window post-take-off. Additionally, sunspot numbers at take-off were included to capture solar cycle states.

[1] LASCO catalogue: https://cdaw.gsfc.nasa.gov/CME_list/
[2] OMNIWeb Plus: https://omniweb.gsfc.nasa.gov/

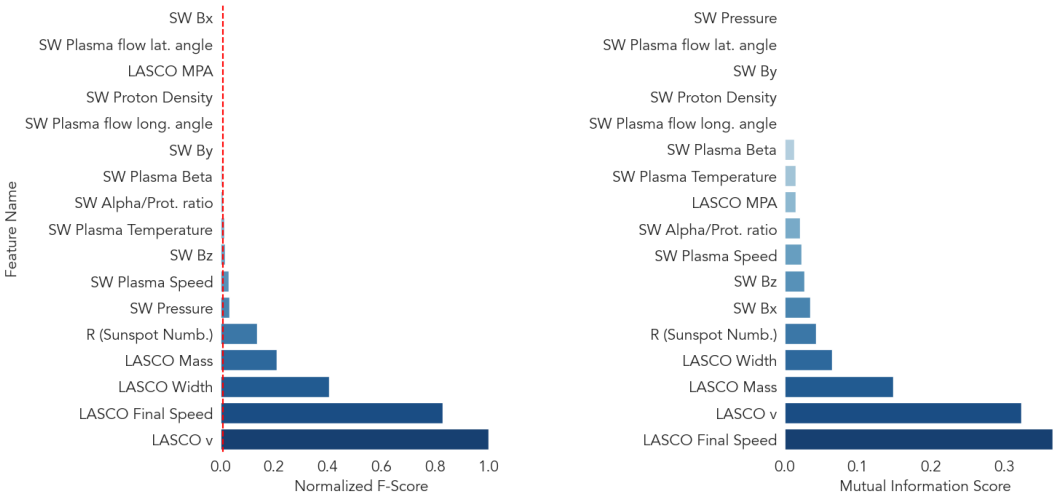


FIGURE 3.1: Barplots of F-score (left) and mutual information score (right) for the regression target (CME transit time).

The relevance of features is assessed using the SelectKBest[3] function from the scikit-learn Python package. This tool supports several feature selection techniques, including the calculation of the Analysis of Variance (ANOVA) F-score and the mutual information (MI) score for both regression (f_regression[4], mutual_info_regression[5]) and classification tasks (f_classif[6], mutual_info_classif)[7]. The F-score evaluates the degree of linear separability between class distributions by analyzing the variance between classes, providing a measure of how well features linearly distinguish targets. In contrast, the MI score captures non-linear dependencies, offering a more comprehensive perspective on feature relevance.

Figures 3.1 and 3.2 illustrate the F-score and MI rankings for features in regression and classification scenarios.

Figure 3.1 reveals that the features most strongly correlated with CME transit time are those related to CME velocity: CME average and final velocity, followed by CME width and mass. Solar wind features also exhibit some relevance, albeit to a lesser degree. The relatively low mutual information values indicate weak non-linear relationships between the data and the target.

[3]SelectKBest documentation: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[4]f_regression documentation: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html#sklearn.feature_selection.f_regression

[5]mutual_info_regression documentation: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html#sklearn.feature_selection.mutual_info_regression

[6]f_classif documentation: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html#sklearn.feature_selection.f_classif

[7]mutual_info_classif documentation: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html#sklearn.feature_selection.mutual_info_classif

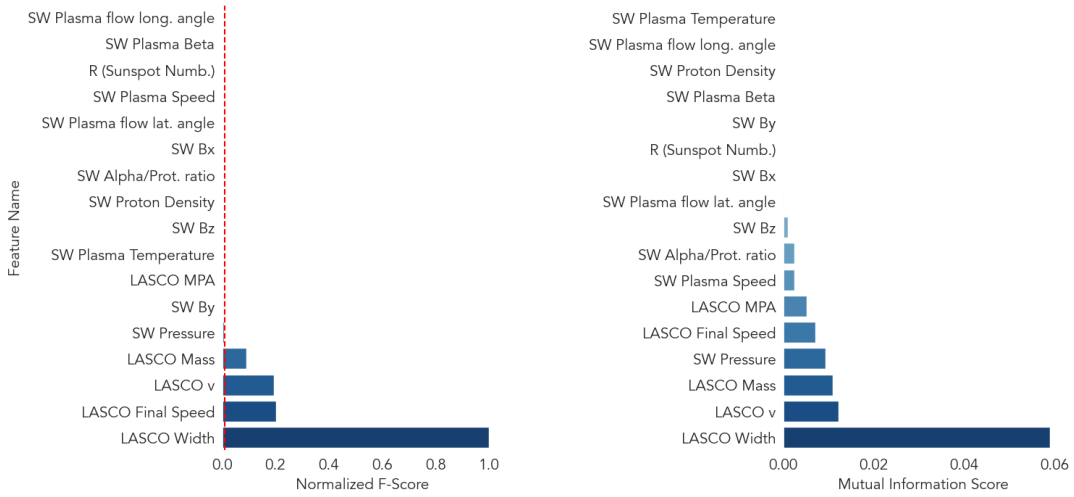


FIGURE 3.2: Barplots of F-score (left) and mutual information core (right) for the classification target.

A similar situation arises with the classification dataset, but in this case, there are only four pertinent features, all concerning the state of the CME at launch. Consequently, we have decided to adopt a consistent approach for feature selection across both cases, in line with the methodology used for CAT-PUMA, which involves eliminating features with a normalised F-Score value below 0.01. Feature selection allows the input space to be resized. For the regression task, the input space now comprises 8 features: the two CME speed features, CME width and mass, solar wind B_z, solar wind plasma temperature, solar wind plasma speed, solar wind pressure, and sunspot number R.

Conversely, the feature space for the classification task only consists of CME average speed, CME final speed, CME width, and CME mass. The ranking of the features for the augmented dataset version is very similar, with the sole difference being that the solar wind Alpha/Proton ratio replaces B_z in the input space.

The set of features selected for regression and classification tasks was designed to capture physically relevant descriptors of CME kinematics and the heliospheric environment. From a physical perspective, parameters such as CME average speed, final speed, and angular width are naturally expected to play a central role in both transit time estimation and hit/miss classification. Faster CMEs propagate more rapidly through interplanetary space, while wider CMEs subtend a larger angular extent, increasing the probability of intersection with Earth's position. CME mass, though observationally more uncertain, reflects the inertia of the ejecta and may influence its susceptibility to drag-induced deceleration. These expectations are broadly supported by the statistical feature relevance scores shown in Figures 3.1 and 3.2. Both the F-score and mutual information rankings assign high importance to CME speed and width, indicating strong discriminative power for both regression and classification targets. CME mass and sunspot number also emerge as moderately informative, suggesting a secondary but non-negligible role in

characterising propagation dynamics and broader heliospheric context. Conversely, features derived from solar wind measurements (e.g., Bz, temperature, dynamic pressure) receive lower relevance scores. This may reflect both physical and methodological factors: solar wind parameters measured at Earth may not accurately represent the conditions encountered by the CME during transit, especially for events with complex trajectories. Furthermore, temporal offsets and averaging may dilute their predictive value in early-stage modelling. Nonetheless, their inclusion provides contextual information that may become more valuable in time-aware or ensemble modelling frameworks. Taken together, the feature rankings align reasonably well with physical intuition, and help validate the chosen input space prior to model training.

Targets and class balance

Two supervised learning tasks were addressed:

1. **Transit Time Prediction:** A regression task predicting the time between CME onset and arrival at Earth.
2. **Earth-Impact Classification:** A binary classification task distinguishing Earth-impacting from non-Earth-impacting CMEs.

To mitigate class imbalance, filters were applied to non-Earth-impacting events. These included removing events with angular widths below 90 degrees or those flagged as 'poor events'. The final dataset comprises two versions:

- *Dataset V.1*: 209 Earth-impacting and 2968 non-Earth-impacting CMEs.
- *Dataset V.2*: 295 Earth-impacting and 3453 non-Earth-impacting CMEs, with missing values imputed using KNN.

3.2 Drag-Based model dataset

Napoletano et al. [2022] compiled a dataset of CMEs by integrating data from the Richardson and Cane CME/ICME list [Richardson and Cane, 2010] and the SOHO-LASCO catalogue [Yashiro et al., 2004][8]. This dataset includes critical information required to solve the DBM equations (1.6, 1.7), which serve as input for the subsequent MCMC algorithm. Some quantities were directly extrapolated from the source lists, while others were derived as part of the analysis in Napoletano et al. [2022]. The dataset encompasses several key parameters, including the of the ICMEs and its associated uncertainty, the VoA of the ICMEs, and the initial velocity (v_0) of the CMEs, along with their corresponding error estimates.

A revised version of this dataset was produced by Mugatwala et al. [2024], and it is publicly available on Zenodo[9]. In this revision, a Monte Carlo

[8]The catalogue is available at https://cdaw.gsfc.nasa.gov/CME_list/
[9]Mugatwala et al. [2024] dataset: https://zenodo.org/record/8063404

approach was employed to invert the DBM equations (1.6, 1.7) analytically, generating a range of possible values for the DBM parameters for each CME. This work introduced two essential advancements: first, the CMEs were clustered based on their affinity to the DBM model, using the acceptance rate from the Monte Carlo inversion to identify the most suitable events for a DBM-based description. Second, the CMEs were categorized as propagating through either fast solar wind conditions (with solar wind speed $w > 500$ km/s) or slow solar wind conditions (with $w < 500$ km/s).

The resulting dataset comprises 213 CME events spanning the period from 1996 to 2018, of which 178 are classified as "slow solar wind events" (slow Solar Wind (SW)), and 32 are categorized as "fast solar wind events" (fast SW).

The dataset includes the ICME velocity of arrival (VoA), initial CME velocity, and associated uncertainties. Events were categorized based on solar wind conditions:

- **Slow solar wind:** $w < 500$ km/s,
- **Fast solar wind:** $w > 500$ km/s.

3.3 Coronal jet dataset

In this study we employ the dataset proposed by Liu et al. [2023] as a baseline; they employed the SAJIA algorithm to full-disk SDO/AIA 304 Åimages from June 1, 2010, to May 31, 2020, with a temporal resolution of six hours. SAJIA yielded 3800 coronal jet candidates. Of these, 1215 were confirmed as true jets by visual inspection. Subsequently, Soós et al. [2024] expanded the analysis by enhancing the temporal resolution to three hours. This refinement led to the detection of an additional 4227 coronal jet candidates within the same timeframe. From these, 1489 were validated as true jets. Overall, the combined efforts resulted in a comprehensive examination of 8027 coronal jet candidates from June 1, 2010, to May 31, 2020. Ultimately, 2704 of these detections were confirmed as true jets.

In their study, Bourgeois et al. [2025] also analyzed full-disk SDO/AIA 304 Å images ranging from 2010 to 2020 but leveraged an MM approach to identify solar structures. Such an approach allows for segmentation of the coronal off-limb structures observable in the full-disk images. The images, pre-processed to remove unwanted chromospheric features, were analyzed using MM to isolate and enhance coronal structures. MM operations such as *erosion* and *dilation* were used in combination to apply a white top-hat transform, which helped in isolating bright coronal features. A fixed threshold was then applied to filter out noise and irrelevant objects, refining the dataset. Such a filtering step is implemented to reduce noise and exclude possible eruptions located too far from the solar disk, which are less likely to be coronal jets. This ensures that our analysis focuses on the most relevant coronal jet candidates. Once the structures were extracted from the images, Bourgeois et al. [2025] computed key properties like area, perimeter, and positional characteristics, such as latitude and longitude, for each structure. The dataset ultimately

comprised 877843 structures. MM proved to be an effective tool for isolating and characterising the complex, dynamic coronal structures observed in the solar corona. More information about the implementation and applications of the MM algorithm can be found in Bourgeois et al. [2025].

In our analysis, we map the structures identified using the MM approach with those detected by the SAJIA algorithm based on their positions on the solar disk. Specifically, we associate each SAJIA jet candidate with the radially closest MM structure. By combining SAJIA and MM datasets, we obtain an MM description of the 8027 structures from Liu et al. [2023]. After the filtering, we obtain a dataset composed of 2667 validated jets (positive events), and 5028 validated as non-jets (negative events).

Figure 3.3 shows the comparison of an exemplary coronal jet detected by SAJIA and the MM approach.

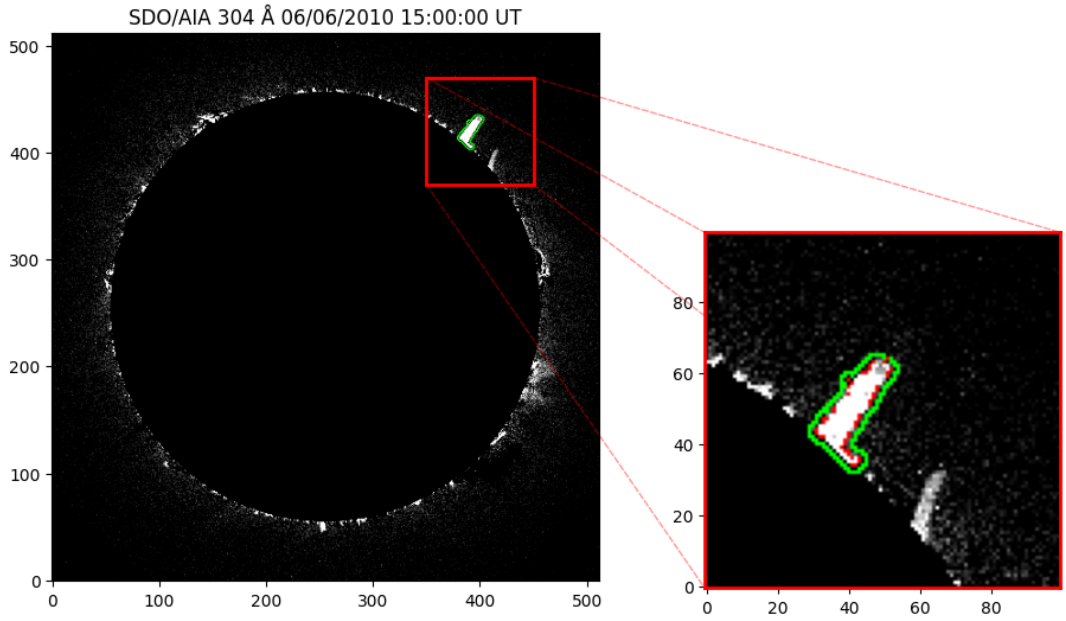


FIGURE 3.3: Comparison of the contouring results from the SAJIA algorithm (red contours) and the MM algorithm (green contours) on the SDO/AIA 304 Å image recorded on 06/06/2010 at 15:00:00 UT.

Figure 3.4 shows the training data obtained from the SAJIA algorithm. A key observation is that the retrieved true coronal jets tend to cluster at high absolute values of latitudes. This suggests that jets are more frequently detected at both high northern and high southern latitudes. Additionally, the number of jet detections is noticeably higher during the early stages of Solar Cycle 24.

This pattern illustrates the spatial and temporal distribution of coronal jets in the training data, highlighting that certain latitudinal regions and phases of the solar cycle are more prone to jet activity. However, this is not representative of the natural behaviour of coronal jets well. Because of the potential bias, we decided not to include latitude and time features in the input space of the model. We employed MM features to encode the descriptions of coronal

jet candidates. Such features were obtained leveraging the DIPlib[10] Python package, which provides access to various morphological metrics such as Feret diameters, radius statistics, convex area, and perimeter. Next, following a feature selection process, we eliminated collinear features to enhance model performance. This process ensures that the remaining features contribute uniquely to the classification task.

At the end of the selection process, the feature space is composed of 17 features, encoding each jet instance. The candidate jets descriptors are the total intensity, the structure area and perimeter, the length-width ratio, the skewness and excess kurtosis of the grey-value image intensities across the object, the Podczeck shape descriptors (square, circle and elongation), the measure of similarity to a circle (circularity), the roundness, the deviation from an elliptic shape (ellipse variance), the bending energy of the structure and finally the position of the closest pixel to the centre of the solar disk defined by the angle and the distance (for detailed information, please refer to the DIPlib documentation).

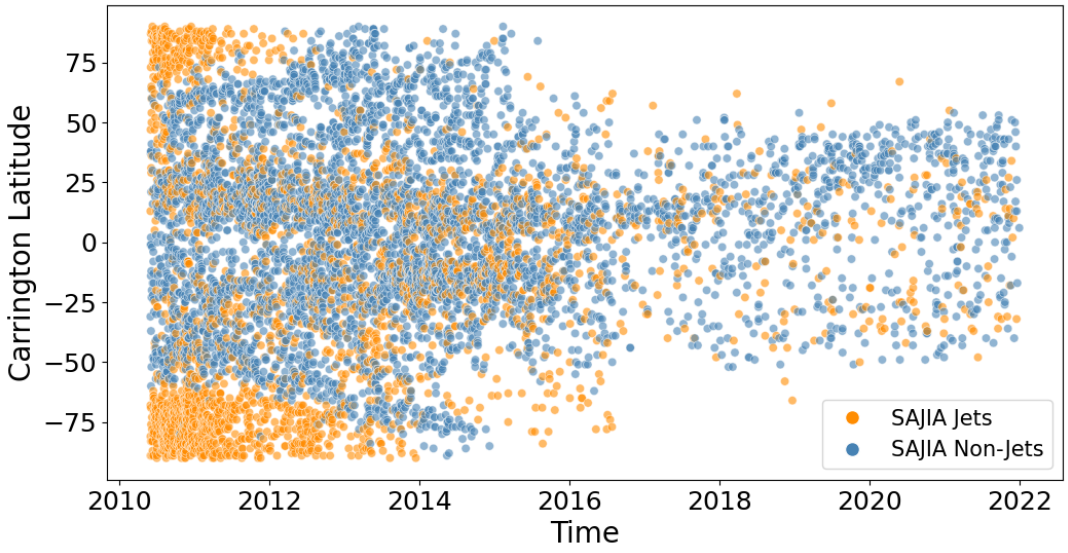


FIGURE 3.4: Scatter-plot of the training data obtained from the SAJIA algorithm. Coronal jets are represented by orange dots, while non-jets are depicted in blue.

The datasets outlined in this chapter form the foundation for the analyses conducted in this thesis. By carefully curating and pre-processing these datasets, as well as implementing feature selection techniques, we ensured that the input data was both representative and optimized for the models employed. The following chapter presents the results derived from applying machine learning and statistical techniques to these datasets, offering insights into the dynamics of coronal mass ejections, their propagation through the heliosphere, and related solar phenomena. In the following chapter, we present the results obtained from these models, showcasing their performance

[10]DIPlib documentation: https://diplib.org/diplib-docs/features.html#size_features_Feret

in addressing the research objectives and providing valuable insights into the phenomena under investigation.

Chapter 4

Results

This chapter presents the main results obtained in this thesis, which are organised into three sections. Each section addresses the findings from different aspects of the research. Section 4.1 focuses on the application and interpretation of the CAT-PUMA model for predicting Earth-impacting CMEs [Chierichini et al., 2024b, published in *The Astrophysical Journal (ApJ)*,)]. Section 4.2 discusses the results related to the revisited P-DBM model and its enhancement using MCMC methods [published in the *Journal of Space Weather and Space Climate (JSWSC)*, Chierichini et al., 2024a]. Finally, section 4.3 presents the outcomes of the dataset expansion and analysis of coronal jets using random forests (submitted to Astronomy & Astrophysics (A&A), which is undergoing revision as of writing).

Together, these sections summarise the core contributions of this thesis to the field of space weather prediction.

4.1 Supervised learning approach to CME arrival modelling

In this section, we describe the results obtained in this work. We set out to train three different ML models and use them for two distinct tasks: regression and classification.

- The regression models provide an answer to the question: How long do CMEs take to reach Earth?
- The classification models generate predictions as to whether a CME will reach Earth or not.

4.1.1 Performance evaluation

We studied various models systematically for both problems under analysis. Each model is optimised to address the relevant ML problem at hand, and then we analyse the performance by comparing different evaluation metrics. For clarity, we will first describe the regression problem and, later, the classification problem.

Regression

The training follows the same steps for all three models:

1. After the feature selection procedure (described in section 3.1) we extract eight relevant features for predicting the transit time; four CME Features and four SW state features.

2. Each model, SVM, random forest and XGBoost is optimised by means of K-fold CV, with $k = 5$ (Sec. 2.1.3).

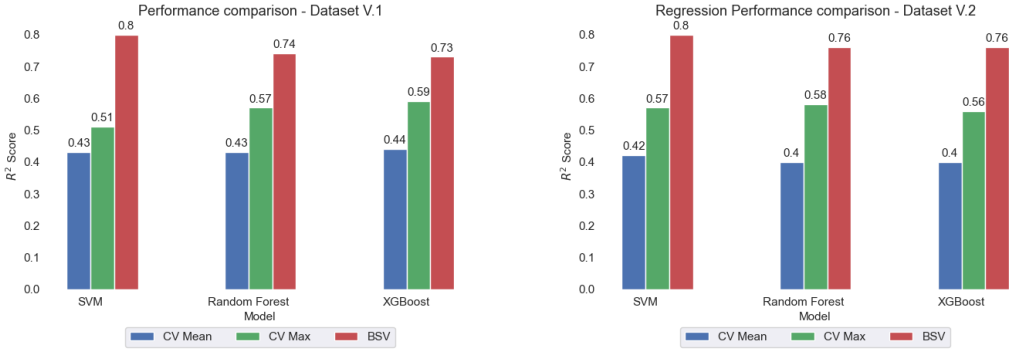
3. Once the optimised models have been obtained, we evaluate their performance through Cross Validation and Best Split Validation (Section 2.1.3), using the R^2 score as a reference metric.

This procedure is applied to both versions of the dataset; the first consists of 209 Earth-impacting CMEs, while the second version contains 295 CMEs, 86 properly imputed as described in Sec. 3.1. Figure 4.1 summarizes the results obtained using different validation methods. We report the average value (blue) and the maximum value (orange) obtained by a 5-Fold CV, as well as the BSV score (green).

Cross-validation is a more conservative method than Best split validation, as mentioned in section 2.1.3, which puts the spotlight on the best Train/Test split. This is evident in the figure; the Best-Split score is the highest for any model-dataset combination. Furthermore, it is essential to point out that best-split validation is less effective for ensemble techniques, returning a lower value than SVM. The reason is probably to be found in the architecture of the models. Ensemble models can better generalise predictions and not fit too closely to the specific Training set used for training. This makes it more difficult to find a Training and Test pair that performs dramatically better than a random split. Nevertheless, the ensemble models also achieve fairly high performance, with a BSV score ranging from 0.73 to 0.76.

The results show a significant difference between the performance according to the BSV and CV scores. The CV Mean scores are similar for all models but are still considerably low compared to the CV Max values. To understand it better, this means that of the five different random Train/Test splits for CV, the most optimistic one returns a considerably higher R^2 score than the average. This is true for both versions of the dataset, underlining the difficulty in characterising a model capable of generalising the regression problem well. We get the best performance from the SVM; the BS validation technique achieves an R^2 score of 0.80, and the related MAE is 7.6 ± 5.2 hours. Although the MAE is higher than in the original version of CAT-PUMA, this result is still reasonably good, considering that the test set includes more events. However, for CV, the MAE is above 10 hours.

It is important to stress this concept; although one can obtain a very high-performing model through BSV , it does not necessarily maintain such high performance on new samples.



(A) (B)

FIGURE 4.1: Performance scores for regression models. Performance comparison by means of CV Mean, CV Max score and Best Split score for dataset V.1 (a) and dataset V.2 (b).

Classification

The second part of this work is devoted to the study of machine learning models capable of predicting whether or not a CME will reach the Earth. The feature selection process (Sec. 3.1) shows that in a classification framework, the features most correlated with the target are only four: LASCO width, final speed, average velocity and mass, all descriptors of the CME at launch.

Again, we opted to test different models on two versions of the dataset. The dataset V.1 includes 2968 CME events, of which 209 are positive (i.e., Earth-impacting CMEs). The augmented version, on the other hand, consists of 3543 CME events, of which 295 are Earth-impacting. For the classification problem, we adopted a more standard validation method. Before training, we divided the dataset into Training (80% of the total) and Test (20% of the total); we optimised and trained the models on the Training set and then evaluated the performance on the Test.

Given the highly unbalanced nature of the problem, it is even more challenging to determine whether and how well a classifier succeeds in solving the problem under analysis. For this reason, we decided to compare several performance evaluation metrics to extrapolate a wider spectrum of information about models' capabilities.

Table 4.1 summarises the results, comparing the values of some relevant metrics to assess the goodness of the classification. There is much information to extrapolate from the results obtained.

First, it is important to emphasise that the performance of the different models is comparable and the score values are generally better for the augmented dataset version (Dataset V.2). Accuracy is higher than 70% in all scenarios. As mentioned earlier (Sec. 2.1.3), however, the accuracy value is not an optimal indicator of the model's goodness because it is affected by the unbalance of the classes. The balanced accuracy value gives a more realistic interpretation of the classifiers' ability to assign the correct class to each instance, never exceeding a value of 65%. In general, the models show

an excellent ability to recognise events in the majority class while lacking Precision for the minority class, resulting in a high False Alarm Ratio (FAR).

Precision is generally very low, reaching a maximum value of 30% for random forest. Nevertheless, the Recall is generally fairly high, indicating the ability of the models to obtain reliable forecasts for non-Earth-Impacting CMEs. It is essential to go into detail on this topic because there is usually a tendency to confuse model performance, which inevitably depends heavily on the type of validation chosen with the actual capabilities of the model.

For the sake of clarity, we provide the confusion matrix for the random forest in figure 4.2. The precision score encodes the following information:

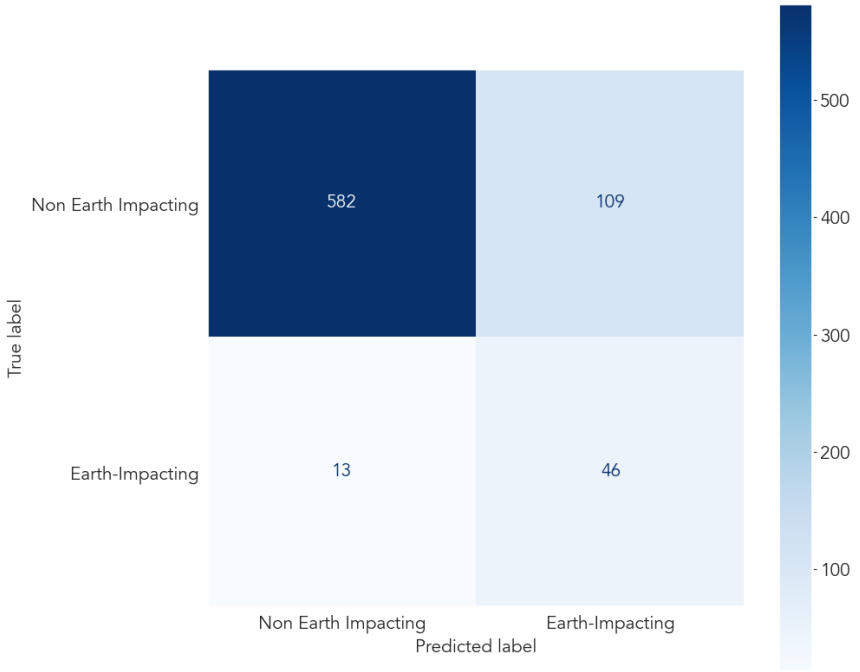


FIGURE 4.2: Confusion matrix for the Test set for the random forest model, trained on the augmented dataset version (Dataset V.2). Matrix entries are TP (bottom right), TN (top left), FP(top left) and FN (bottom left).

among 155 events predicted as Earth-Impacting, only 46 are correctly classified. Low precision directly implies a high false alarm ratio. Despite this, the model still shows potential for operational application because of the high Recall. In fact, of 686 events labelled as Earth-Impacting, only 13 are predicted incorrectly.

4.1.2 Interpretation of results

One of the main criticisms levelled at prediction tools based on machine learning algorithms is that it is difficult to judge their actual capabilities and limitations because there is often no way of getting a sense of the process that drives the models to produce a specific prediction. In addition, hard-to-interpret models such as deep neural networks and gradient-boosting

TABLE 4.1: Comparison of evaluation metrics for SVM, random forest, and XGBoost models across two dataset versions (V.1 and V.2). Metrics include Accuracy, Precision, Recall, Balanced Accuracy, and False Alarm Ratio. The bold values highlight the best performance within each metric across the dataset versions.

Metric	SVM	random forest	XGBoost
Accuracy	0.76 | 0.77	0.82 | **0.84**	0.73 | 0.78
Precision	0.19 | 0.24	0.24 | **0.30**	0.18 | 0.24
Recall	0.71 | **0.85**	0.67 | 0.78	0.79 | 0.81
Balanced Accuracy	0.58 | 0.61	0.60 | **0.64**	0.58 | 0.61
False Alarm Ratio	0.81 | 0.76	0.76 | **0.70**	0.81 | 0.76
		Dataset V.1 | Dataset V.2	

machines are increasingly efficient and now outperform, in most cases, linear models that are typically easier to interpret. The main consequence of the lack of interpretation is distrust in the model. Can I actually trust a model that I do not fully understand?

The subject of interpretation has been widely discussed in recent years, and various methods have emerged to better understand the results obtained by artificial intelligence. Local explanation methods aim to assess the influence of input variables/features on a specific prediction/output. In this paper, we want to exploit one of these tools, called *Shapley values* [Lundberg and Lee, 2017], to gain more insights into model decisions.

SHAP is a model-agnostic local explanation method originated in the field of game theory to determine the payouts of players depending on their contribution to the total payout [Aas et al., 2021]. In an Artificial Intelligence (AI) explanation setting, this method is used to calculate the contribution of each feature to the final output. In particular, this technique allows us to decompose the output of a model $f(\bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ is a specific feature vector, into the sum of the contributions ϕ of each feature:

$$f(\bar{\mathbf{x}}) = \phi_0 + \sum_{i=1}^{F} \phi_i. \tag{4.1}$$

Considering a set of F features and a subset $S \subseteq F = \{1, ..., F\}$ consisting of $|S|$ features, the Shapley value related to feature j can be expressed as:

$$\phi_j(v) = \phi_j = \sum_{S \subseteq F} \frac{|S|!(F - |S| - 1)!}{F!}(c(S \cup \{j\}) - c(S)), \quad j = 1, ..., F, \tag{4.2}$$

where $c(S)$ is the contribution function that maps subsets of features to the contribution they have on the prediction. Such function is typically the expected output of the model, conditional on the feature vector $\mathbf{x}_s$:

$$c(S) = E[f(\mathbf{x})|\mathbf{x}_S = \bar{\mathbf{x}}]. \tag{4.3}$$

In essence, the Shapley Values determine the difference in the contribution that feature j brings to the prediction if included in a specific subset S and average this over every possible combination of possible subsets S of features. (in terms of the contribution function: c(subset S including feature j)$-$ c(subset S without feature j)

In this work, we used the python package SHAP[1] to apply the theory of Shapley values to the predictions made for CMEs and try to obtain some more information on the feature space of the CAT-PUMA framework.

The SHAP visual tools help to quantify the contribution of each input feature to the model's output for a given prediction. The summary plot aggregates the impact of features across all data points, highlighting which features most significantly influenced predictions across the dataset. The decision plot, in contrast, provides a cumulative view of how individual features push predictions higher or lower for each specific instance. The waterfall plot is particularly useful for understanding the model's reasoning for a single prediction. It visualises how the base value (i.e., the average model output over the training dataset) is adjusted by each feature's contribution to reach the final predicted value. Features that increase the prediction are shown in red, while those that decrease it are in blue. This allows a transparent and interpretable breakdown of model decisions, especially for complex ensembles like random forests or gradient-boosted trees.

Since we tested different machine learning models, we decided to deal in more detail with the cases where performance is highest to see if there are patterns that characterise the best-performing models. As with the description of the results, we will start by treating the regression case and then discuss the classification task.

Regression

For the regression case, we considered the SVM model trained on the dataset V.1. One of the main tools offered by the SHAP algorithm is the summary plot shown in Figure 4.3 (A), which shows for each feature the SHAP values of all instances in the training set. This plot contains a lot of information about the predictions made by the model, so we try to break down the main ones; first of all, the features on the y-axis are ordered in ascending order (from bottom to top) according to the average contribution they have on the predictions. This means that, according to SHAP, the feature with the greatest influence on the predictor output is the average speed of CMEs, followed by Angular width, final speed and sunspot number R, while the least influential features are the SW Speed Temperature and pressure.

In addition, SHAP values are typically higher for low feature values and lower (negative range) high feature values; this is true for all features, especially speed features, except for SW B_z. In practice, very high feature values tend to push the model predictions towards lower transit time. Trivially, if the speed of the CME is very high, the model will tend to opt for low transit time estimates. Another convenient way of obtaining information on the

[1]SHAP documentation: https://shap.readthedocs.io/en/latest/

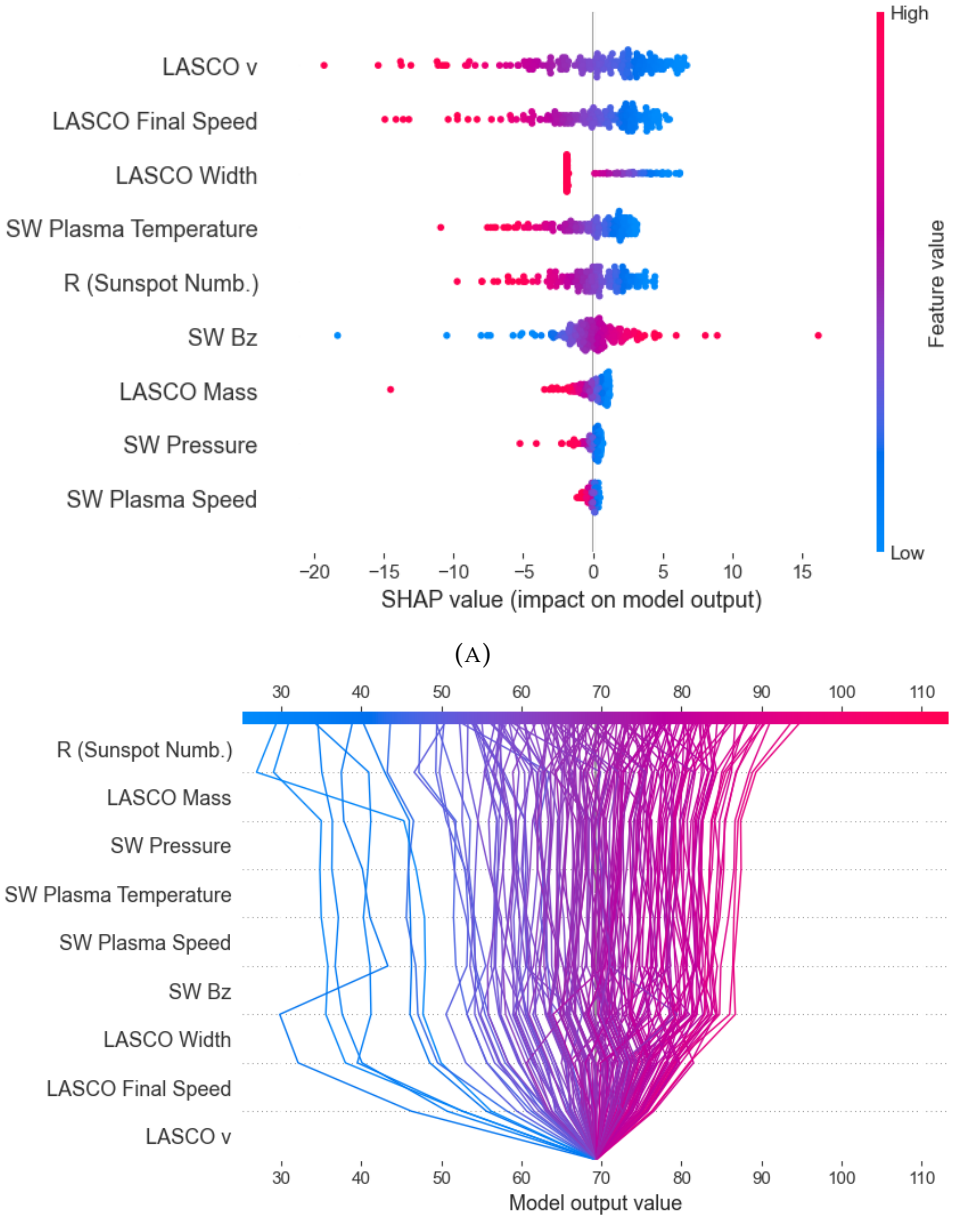


FIGURE 4.3: (A) SHAP Summary plot for the training set. The y-axis ranks the features sorted from the most (top) to least (bottom) important. The x-axis depicts the SHAP value. Each point refers to a specific instance of the training set, pointing out the related SHAP value associated with a value of a certain feature. The colour bar displays whether the feature value is high (pink) or low (blue). (B) SHAP Decision lot for the training set. This plot shows the decision path for each instance in the training set. Each line shows each feature's contribution (x-axis) to the final output of the model. The colour depends on the magnitude of the output and ranges from blue for lower output values to red for higher ones.

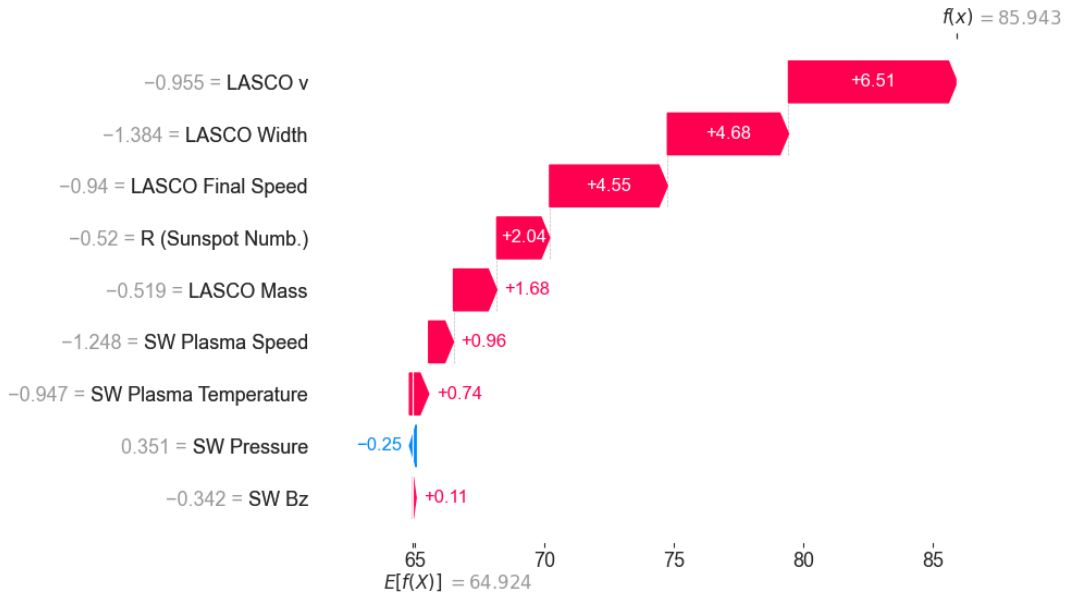
model's decision-making process is the decision plot (Figure 4.3 (B)).

This tool helps visualise the decision path that the model takes for each instance. For each instance in the training set, the graph shows the contribution each feature has on the final output. The paths are clustered by similarity, which allows similar decision patterns to be identified. Two different macro-patterns can be distinguished; the first relates to most instances and mainly involves output values of more than 50 hours, while the second refers to low transit time predictions. For all instances, the LASCO Speed- and Width-related features direct the prediction the most. All other features have a lower impact, and at the top of the cascade, the sunspot number R produces the push towards the final output of the model. The instances associated with lower predictions (transit time < 50) appear to be largely conditioned by the velocity value of the CMEs at launch time; this suggests that if the initial velocity of the CMEs is very high, the model is likely to generate lower transit time predictions. Furthermore, the decision pattern for low transit time predictions appears less stable, there are a couple of cases where B_z and LASCO Mass values push the predictions considerably towards higher or lower output, respectively.

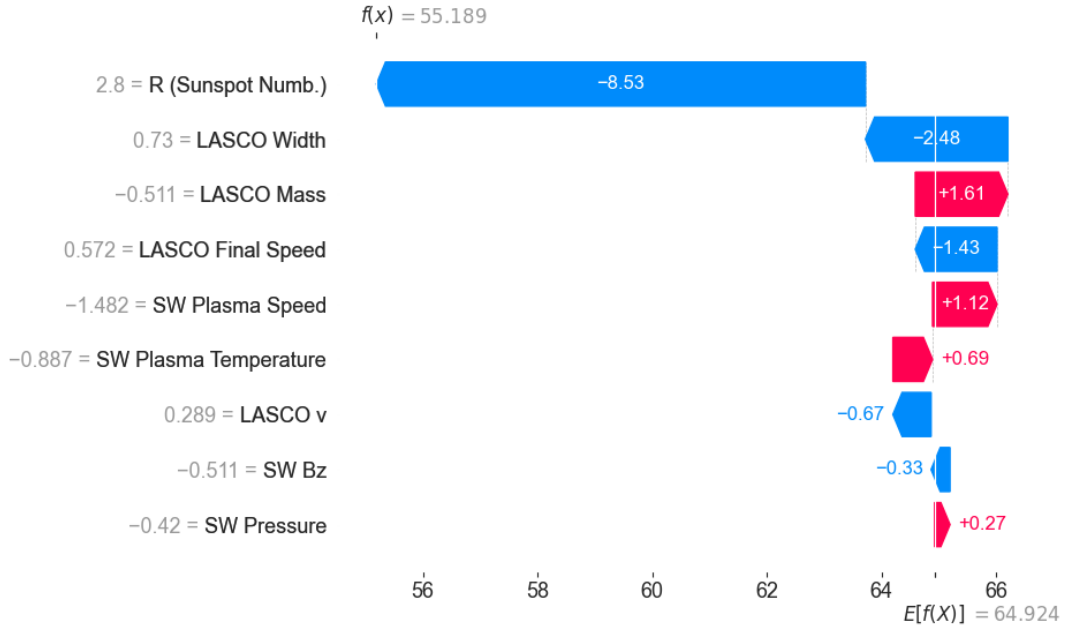
This is interesting because, in fact, there are relatively few examples of CMEs associated with a very low transit time (<40 hrs); this might suggest that due to the few examples available, the model appears to rely more on speed features to make decisions about lower outputs. This is because the correlation between the transit time and the speed of the CMEs is higher, and it is, therefore, easier to establish a relationship with the few examples available. Moreover, SHAP, being a local technique, is valuable for inspecting decisions on individual instances. Waterfall plots of the instances with the highest and lowest prediction error are shown in the figure. Such plots can clearly and compactly display the relative contributions of the different features in order of importance. Waterfall plots of the instances with the highest and lowest prediction error are shown in Fig. 4.4. Such plots can clearly and compactly display the relative contributions of the different features in order of importance. The least performing instance has a recorded arrival time of 108 hours. In Figure 4.4 (A), we see how almost all the features push the output towards very high transit time values but fail to reach the actual value, which is still very high compared to the average value. This effect is probably still due to our poor representation of rare events in the training set, as we do not have many examples of such slow CMEs in our dataset.

In contrast, the best-performing CME is associated with a transit time that is much closer to the mean value. Figure 4.4 (B) shows that almost all features hold the prediction value close to the base value. The sunspot number R has the most significant contribution by pushing the prediction very close to the actual value, resulting in an error of only 0.5 hours.

Let us now move on to the analysis of the classification task.



(A)



(B)

FIGURE 4.4: Waterfall plot related to the best (A) and worst (B) performing CME.The plot shows the relative contribution of each feature to the model's prediction $f(x)$, starting from the base value $E[f(x)]$. The x-axis shows the features and their value (scaled for training), while the x-axis represents the transit time. The arrows display the SHAP value associated with each feature, coloured red if positive and blue if negative.

Classification

In this section, we delve into the decision-making process leading to the predictions in the classification task; in particular, it is interesting to exploit the SHAP values to find insights as to why the FAR remains so high. For this purpose, we analyse the predictions made on the test set by the best-performing model, the random forest trained on the V.2 dataset. The model outputs are values between 0 and 1, and instances are associated with the positive or negative class by identifying a threshold value, usually 0.5; thus, samples with an output greater than the threshold value are associated with the positive class. Otherwise, the prediction is negative. The output score also indicates how confident the model is in making decisions. The closer the output value is to the threshold value, the more uncertain the decision possibly is. Figure 4.5 (A) shows the classification confidence for the CMEs in the test set; the histogram suggests that for most of the misclassified CMEs, the model decision was made with confidence of less than 0.7; in contrast, correctly classified CMEs typically have very high confidence, in most cases greater than 0.8. This suggests that despite its high FAR , the model is relatively confident when making a correct decision, while it is generally less secure when it makes incorrect predictions. This result is reassuring because it suggests that the model learns the difference between Earth-impacting and non-Earth-impacting CMEs.

The SHAP method allows the model's decision-making process to be analysed instance by instance. Figure 4.5 (B) shows the decision plot for the misclassified test set events. The decision plot highlights some interesting aspects. First of all, we notice two main decision patterns; in blue are the CMEs assigned to the negative class and in red those assigned to the positive class. There are also some instances in which the model associates an output value very close to the base value, i.e. close to the threshold value. The latter shows a more uncertain decision pattern, with some features pushing them towards higher values while others lowering their output value; the result is an output that settles close to the threshold value. Instances with an output greater than the threshold value, thus assigned to the positive class, contribute to the high FAR . For those instances, the graph shows that the feature that most influences the decision is the LASCO width, which pushes the prediction towards high values. However, the other features tend to lower the output value by pushing back the output, making the model's decision less secure.

This is interesting because it suggests that in these cases, the model is principally 'confused' by the value of the width of the CMEs; most of the misclassified events are Halo CMEs (LASCO width = 360 degrees). Considering the instances incorrectly assigned to the negative class, although the number of such events is very low, the decision plot suggests that the model is generally relatively confident in the choice since almost all features push the output towards values close to zero, although there are few misclassified instances in this case.

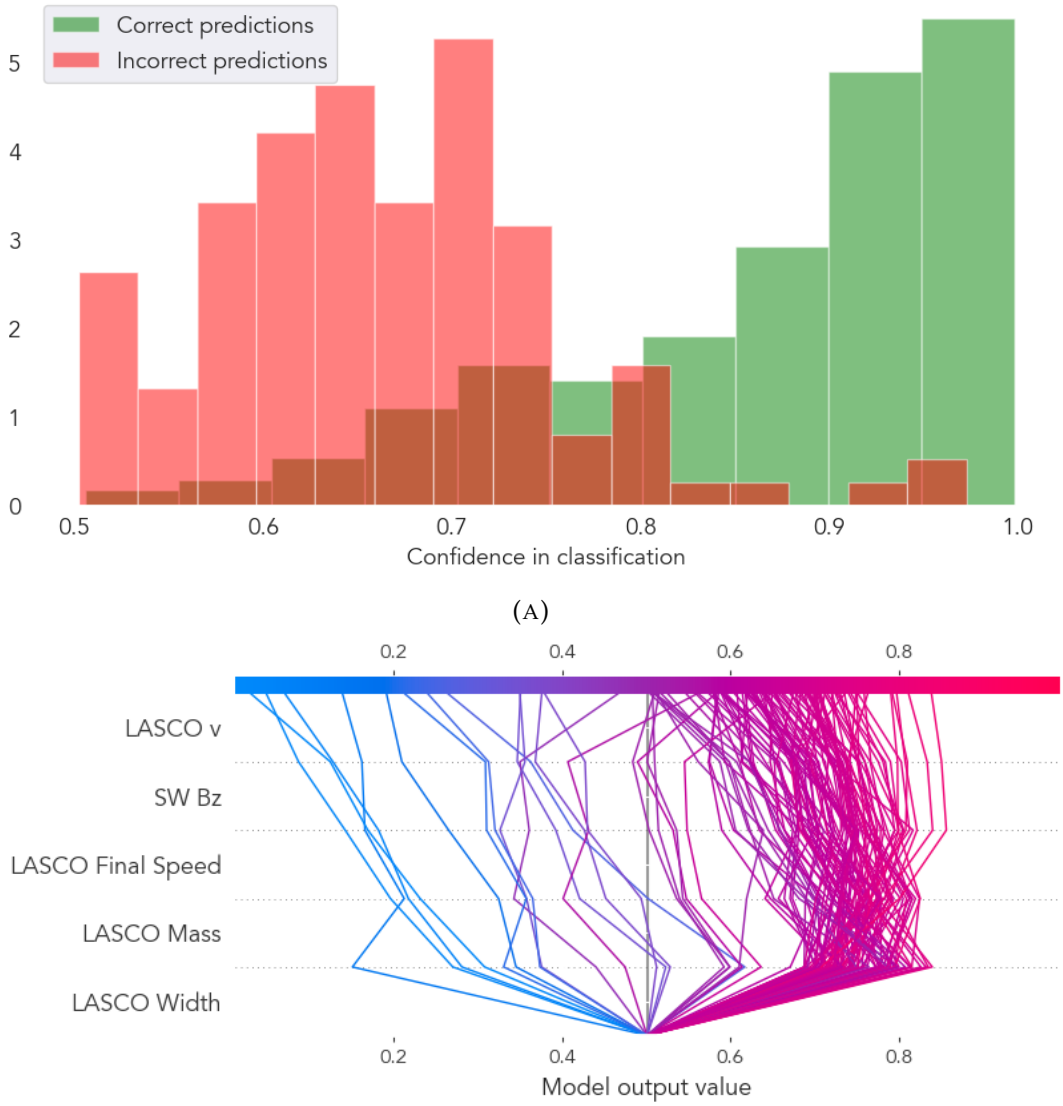


FIGURE 4.5: Interpretability plots for the classification Task. The graphs refer to the Test set for the best-performing classifier (random forest trained on Dataset V.2). (A) Histogram of the classification confidence distribution. Red highlights the misclassified instances, while green highlights the correct predictions. (B) Decision plot for the misclassified instances. This plot shows the decision patterns; the colour bar indicates the magnitude of the output; in blue, those instances for which the model returns values close to zero (assigned to the negative class) are highlighted. In red are those associated with the positive class. Examples related to values close to the base value (i.e. the threshold value) are purple.

4.2 A Bayesian approach to the drag-based modelling of ICMEs

In this section, we present the results of our analysis, including convergence diagnostics, the statistical properties of the parameter distributions, and the model's forecasting performance. For clarity, the discussion is divided into two subsections: the ensemble approach and the individual approach.

4.2.1 Ensemble approach

The objective of the ensemble approach is to derive the PDFs of the DBM parameters γ and w for two specific categories of CME events: those associated with slow solar wind (slow ensemble) and those with fast solar wind (fast ensemble). To ensure robust DBM descriptions, we include only CMEs classified as "Nice fits" by Mugatwala et al. [2024], which are deemed most suitable for DBM analysis. This selection mitigates the risk of convergence issues in the algorithm's posterior PDFs by excluding unsuitable events. The slow ensemble consists of 87 CMEs, while the fast ensemble includes 15 CMEs.

The MCMC algorithm utilizes the available CME data to solve the DBM equations, with the prior distributions encoding existing knowledge about the parameters. We perform the following analysis to assess the convergence and stability of the resulting posterior PDFs. We generate four subsets for both the slow and fast ensembles by randomly sampling 80% Each subset is subjected to four independent MCMC chains, each starting from a different point in the parameter space, with 10,000 iterations per chain. This number of iterations strikes a balance between computational efficiency and the acceptance rate of the resulting parameter distributions. Thus, 10,000 parameter samples are generated for each subset, resulting in a total of 40,000 samples per subset.

Figure 4.6 illustrates the algorithm's evolution for the slow ensemble. Despite the chains starting from different initial conditions, they converge to the same region of the parameter space within each subset, indicating consistent sampling of the posterior distribution. Out of the 10,000 samples generated from each of the four chains, the first 900 samples are discarded as part of the burn-in phase. To reduce autocorrelation, the chains are thinned by retaining one sample every 30 iterations based on the estimated autocorrelation time. As a result, each subset contains 1,256 samples after burn-in and thinning for both the fast and slow solar wind cases. This procedure is applied uniformly across all four subsets.

Figure 4.7 presents the histograms of the marginal distributions for γ (left) and w (right) derived from the four subsets for the fast (top) and slow (bottom) solar wind cases. Additionally, Figure 4.8 displays the cumulative distribution functions (CDFs) for the same subsets.

The PSRF , as discussed in Sec. 2.2.5, is used to assess convergence by measuring the ratio of intra-chain variance to inter-chain variance. A PSRF value close to one indicates that the chains are effectively sampling the same region of the parameter space, confirming convergence. The PSRF scores (reported in Figure 4.7) confirm the convergence of the chains across all cases.

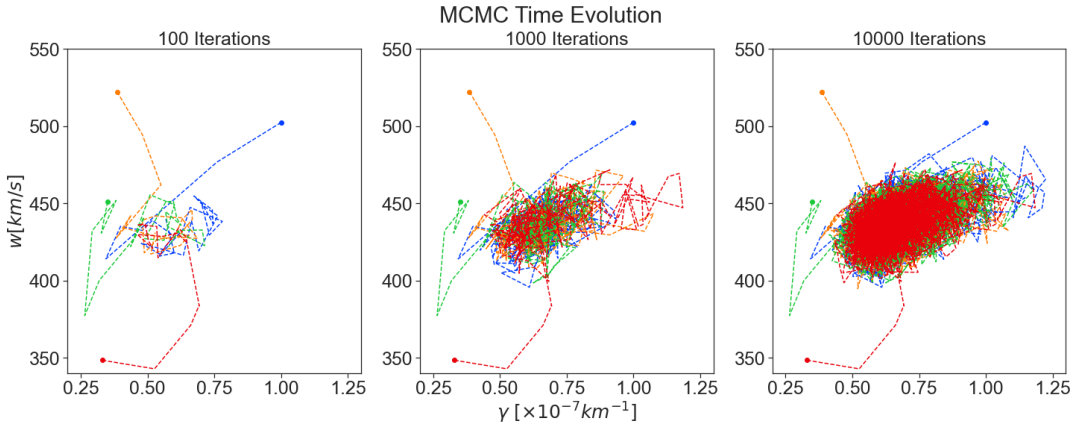


FIGURE 4.6: MCMC evolution plot illustrating the progression of the algorithm for the slow ensemble across three stages: 100, 1000, and 10,000 iterations. The four chains' initial points (depicted as dots) are drawn from an over-dispersed distribution relative to the target density. With the progression of iterations, all chains converge toward the same region of the parameter space defined by the DBM parameters γ and w.

Additionally, the PDFs of the different subsets exhibit highly similar mean values. The standard deviation of the average values of the PDFs is close to zero, as shown in Figure 4.7, indicating that the algorithm remains stable despite slight variations in the dataset.

These results demonstrate the algorithm's robustness in terms of both convergence and stability. We can, therefore, conclude that all the extracted samples are drawn from the same stationary posterior distribution, which successfully distinguishes the fast case from the slow case. Figure 4.9 presents the joint and marginal PDFs of γ and w. In the fast SW case, the posterior PDF of the solar wind speed (w) exhibits an average value of 600 km/s, while in the slow case, the average value is 430 km/s. Notably, in the fast SW case, w values do not drop below 500 km/s, whereas in the slow case, the highest value remains below 480 km/s. These findings align with the expected behaviour of CMEs propagating in slow and fast solar wind conditions.

However, the marginal distributions of the drag parameter (γ) reveal significant differences. The drag parameter, which models the interaction between the CME and the solar wind, tends to be larger in the fast SW ensemble compared to the slow SW ensemble, as shown in Figure 4.9 (lower left). Additionally, a slight correlation between w and γ is observed in the slow SW case (Figure 4.9, upper right), where an increase in w corresponds to an increase in γ. The dispersion around the mean values also differs between the two ensembles. In the slow SW case, the MCMC algorithm produces values that cluster more tightly around the mean, leading to a smaller standard deviation. This tighter distribution can be attributed to the larger size of the slow SW ensemble compared to the fast SW ensemble. A key constraint in the ensemble approach is that new samples (γ, w) are only accepted if they solve the DBM equations for all CMEs in the ensemble. This constraint makes the

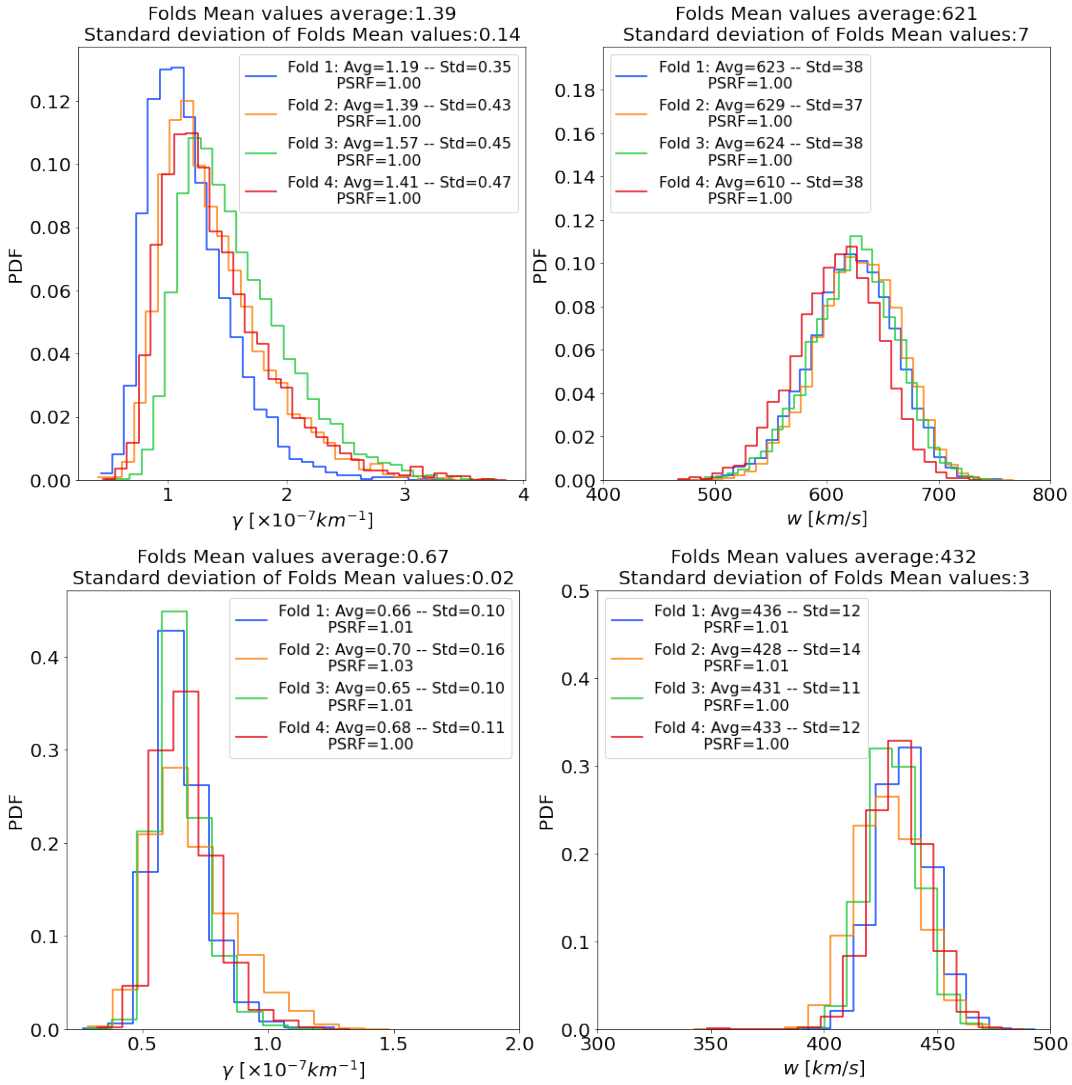


FIGURE 4.7: Probability distribution functions for solar wind speed w and drag parameter γ for fast (top) and slow (bottom) CME obtained leveraging four different folds of the dataset. The legend reports the mean value (avg), the standard deviation (std) and the PSRF score of the folds.

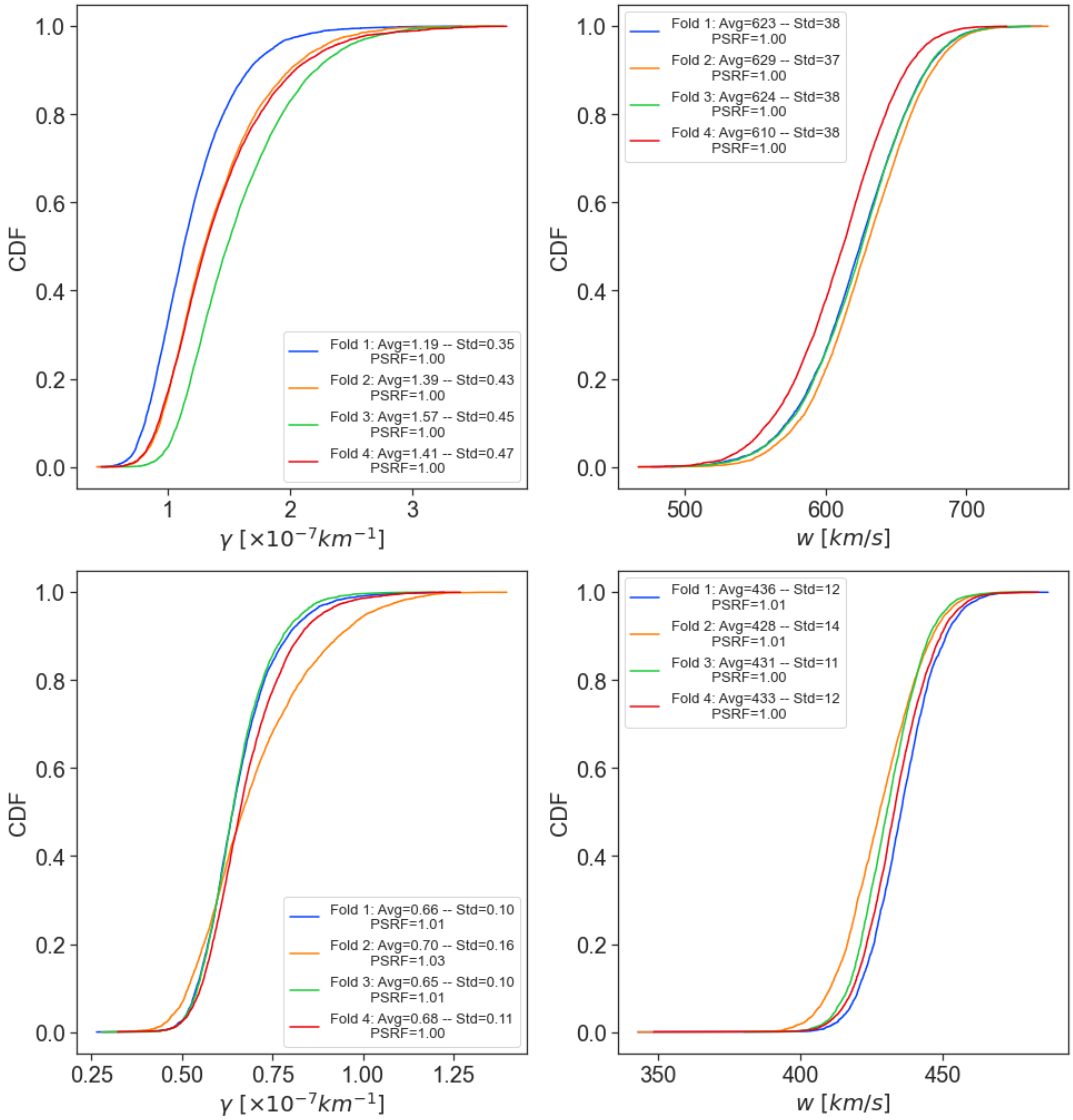


FIGURE 4.8: Cumulative distribution functions for solar wind speed w and drag parameter γ for fast (top) and slow (bottom) CME obtained leveraging four different folds of the dataset. The legend reports the mean value (avg), the standard deviation (std) and the PSRF score of the folds.

slow SW case more conservative, as the accepted samples must account for a wider range of events compared to the fast SW case.

The resulting PDFs represent the parameter values that best describe the average behaviour of CMEs in the ensemble. However, these values may not necessarily represent the optimal pair of parameters (γ, w) for all individual CMEs in the group. In this framework, the dynamics of each CME are modelled probabilistically, with the DBM parameters (γ, w) treated as distributions rather than fixed values. Given the natural variability in solar wind speed due to the solar cycle, solar rotation, and the different sources of the wind on the Sun, the PDFs for the DBM parameters are expected to vary across CMEs. Thus, the ensemble approach focuses on identifying parameter samples that best fit the average behaviour of CMEs in the DBM framework. In the following section, we will present the results obtained using these PDFs to forecast CME transit times.

Validation: transit time forecasting

One of the primary applications of the P-DBM framework is to forecast the transit time and impact speed of CMEs, along with their associated uncertainty. This is achieved within a probabilistic framework by leveraging the estimated PDFs of the DBM parameters (γ, w) to generate an ensemble of predictions for the transit time. The mean of the predictions serves as the estimated transit time, while the standard deviation provides the associated uncertainty.

In this study, we evaluate the forecasting capabilities of the PDFs derived from the P-DBM framework using a cross-validation technique. The dataset is divided into four training and test folds, with 80% of the events randomly selected for training and the remaining 20% reserved for testing. The training folds correspond to the four subsets described in the previous section, and the PDFs are generated from the training set. The forecasting performance is then evaluated on the test set.

For the slow case, the four training subsets consist of 68 events, while the test subsets include 17 events. In the fast case, the training subsets consist of 12 events, and the test subsets contain 3 events. In total, we evaluate 68 slow and 12 fast test events. This cross-validation approach ensures robustness in the performance evaluation and provides a sufficiently large test sample to assess P-DBM's forecasting capabilities. Using the P-DBM framework, we generate distributions of predicted transit times rather than single-point estimates. The mean value of this distribution is treated as the estimated transit time ($\hat{T}$) for each CME, and we expect the true transit time (T) to fall within the 1σ confidence interval in approximately 68% of cases.

The forecasting results are summarized in Figure 4.10. The DBM achieves prediction performance in line with the literature, with MAEs of approximately 10 hours for the slow case and 7 hours for the fast case and standard deviations of 3.4 hours and 0.8 hours, respectively (Figure 4.10, left). The forecast residuals (Figure 4.10, left) show minimal bias in the fast case (-0.9 hours) and a slight underestimation in the slow case (-4.55 hours) in terms of MAE. However, from a probabilistic perspective, the performance of

P-DBM is relatively low, as the true transit times fall within the 1σ confidence intervals in fewer than 68% of cases for both the slow and fast ensembles. We attribute this inconsistency to the structure of the inference method, which imposes strict constraints on the DBM parameter values, as discussed in Section 4.2.1. These constraints result in narrow posterior PDFs and the acceptance of samples with high-likelihood errors.

In the next section, we will discuss the individual approach to modelling CME transit times.

4.2.2 Individual approach

The ensemble approach provides PDFs of the DBM parameters for a group of CMEs. In contrast, the individual approach seeks to further explore the potential of the MCMC algorithm by generating specific P-DBM descriptions for each CME event in the dataset. While the overall structure of the algorithm is similar to that of the ensemble approach, several key distinctions are introduced.

First, the input data now pertains to individual CMEs, with the aim of producing output specific to each event. The resulting PDFs describe the DBM parameters for each CME, without the constraints imposed by fitting all CMEs in an ensemble. A notable difference is the introduction of the CME initial velocity (v_0) as a free parameter in the MCMC algorithm to account for the heterogeneous error distribution of v_0 in the dataset, which can otherwise hinder convergence. We fix the heliospheric distance to 1 AU to keep the degrees of freedom to a limited number. Achieving convergence in the individual approach is more challenging compared to the ensemble case, as the dynamics of each CME event may be described by different DBM parameters. Thus, we adopt weakly informative prior PDFs. For example, a broad Gaussian distribution is used for the solar wind speed w, with a mean of 400 km/s and a standard deviation of 200 km/s. The prior for v_0 is centred on the dataset's values with a standard deviation of 200 km/s, while a log-normal PDF is chosen for the drag parameter γ. These priors guide the sampling process toward regions of the parameter space close to the most likely values.

The convergence study remains unchanged, with four MCMC chains initialized with different starting values for each CME. The PSRF score is recorded, and after the burn-in phase, the chains are thinned as before. This approach allows for the independent investigation of each CME, with PDFs generated for the DBM parameters of every event. For each CME, we record statistical indicators such as the mean and standard deviation of the samples, chain convergence, and the algorithm's acceptance rate. Since the algorithm is applied individually, there is no predefined distinction between CMEs categorized as slow or fast. Instead, we define slow and fast ensemble PDFs by concatenating the samples of CMEs labelled slow and fast, respectively, according to Mugatwala et al. [2024]. In essence, the ensemble PDFs are constructed by aggregating the individual PDFs of the relevant CMEs.

Figure 4.11 displays the histograms of the marginal PDFs for the DBM parameters in the slow and fast ensembles obtained using the individual approach. The distribution of w values for the fast ensemble is noisier compared to the slow ensemble, likely due to the smaller number and greater heterogeneity of fast events. Nonetheless, the results are consistent with those obtained from the ensemble approach, though the marginal distributions are broader. The mean w values are 415 km/s for the slow case and 514 km/s for the fast case, while the average γ values are $0.82 \times 10^{-7}\text{km}^{-1}$ and $1.04 \times 10^{-7}\text{km}^{-1}$, respectively.

These findings indicate that even in the individual approach, the algorithm tends to prefer $w < 500$ km/s for slow CMEs and $w > 500$ km/s for fast CMEs. Additionally, γ assumes higher values in the fast case, with a distribution that exhibits a longer tail. It is important to note that the algorithm does not achieve convergence for all CME events. Some events exhibit non-robust convergence, as the PSRF score and acceptance rate data indicates. To focus on the most robust results, we selected events with a PSRF score of less than 1.05 for all free parameters and an acceptance rate above 5%. Of the 213 CMEs in the dataset, 117 meet these convergence criteria. Among the 102 events categorized as "Nice Fits" by Mugatwala et al. [2024], 64 demonstrate good convergence.

A further noteworthy observation is the inconsistency between the average values of the PDFs obtained through MCMC and the fast and slow labels provided by Mugatwala et al. [2024].

Figure 4.12 shows that some events labelled as slow exhibit MCMC PDFs with mean solar wind speeds exceeding 500 km/s. Additionally, most events with high solar wind speeds occur during the ascending or descending phases of the solar cycle. The resulting labelling scheme for CMEs, based on the average values of the PDFs obtained through MCMC, has been adopted. Of the 117 well-converged CMEs, 90 have an average solar wind speed of $w < 500$ km/s and are labelled as MCMC Slow, while 27 have $w > 500$ km/s and are labelled as MCMC Fast. Figure 4.13 shows the PDFs of the new ensembles alongside those from Mugatwala et al. [2024] (in grey). The PDFs remain similar, spanning approximately the same range of values. The distribution of w for the new slow ensemble shifts toward lower values, with an average of 400 km/s. In contrast, the new fast ensemble gathers higher w samples, with an average of 580 km/s. A shift is also observed in the distributions of the drag parameter γ. The mean values for both the MCMC and Mugatwala et al. [2024] ensembles are higher than before, and the gap between them widens. Notably, the tail of the distribution for the new fast ensemble is thicker and longer. Despite a larger sample size, the fast ensemble's distribution remains somewhat noisy, particularly for w. Finally, the individual approach PDFs are used to test the forecasting capability of CME arrival times.

Validation: transit time forecasting

In this section, we present the results of CME transit time forecasting using the individual approach with P-DBM. The individual approach enables us to generate specific PDFs for each CME event in the dataset, which can then be aggregated to define PDFs for ensembles of CMEs with common characteristics.

We define two versions of the PDFs for the slow and fast ensembles. In the first version, the PDFs are constructed using the labels from Mugatwala et al. [2024], resulting in 87 slow events and 15 fast events. In the second version, we expand the dataset to include all CMEs that meet the convergence criteria, yielding 90 slow events and 27 fast events. To evaluate forecasting performance, we employ a 4-fold CV, similar to that used in the ensemble approach. The dataset is divided into four sub-ensembles: three for training and validation to define the PDFs and one as a test set for evaluation. For the first version, each training set contains 68 slow events and 12 fast events, while each test set contains 17 slow events and 3 fast events. In the second version, the training sets consist of 72 slow events and 22 fast events, and the test sets contain 18 slow events and 5 fast events.

The forecasting results for both versions are summarized in Figure 4.14.

The graphs in Figure 4.14 display the forecasting results obtained using the individual approach with P-DBM. The upper graphs correspond to the ensembles defined by the labelling from Mugatwala et al. [2024], while the lower graphs represent the ensembles relabeled using the MCMC approach. Overall, the forecasting performance of the individual approach is consistent with that of the ensemble approach, with comparable MAE values indicating similar levels of accuracy. In the first version of the PDFs, the slow ensemble shows a slightly lower average error, while the fast ensemble has a higher average error.

Although the model performs well from a probabilistic perspective, the wide error bars associated with the transit time estimates reflect the broadness of the PDFs. Similar to the ensemble approach, the model tends to underestimate transit times. However, the results from the MCMC relabeled ensemble PDFs are less promising, with higher MAE values indicating larger errors. The overestimation of fast CMEs is particularly noticeable. This discrepancy is further reflected in the probabilistic performance, where the 1σ confidence intervals are not consistently met. In contrast, the forecasting performance for the slow ensemble remains satisfactory, likely due to the larger size of the test set.

4.3 Coronal jet identification with machine learning

This section is devoted to the identification of coronal jets using machine learning techniques. The focus is on the development of a comprehensive dataset of coronal jets by combining results from multiple identification

TABLE 4.2: Confusion matrix for the random forest classifier

Actual / Predicted	Non-Jet	Jet
Non-Jet	835	187
Jet	182	346

methods, such as SAJIA and MM. By employing the random forests algorithm, the analysis provides a robust framework for expanding the existing dataset and improving the accuracy of jet identification. These efforts aim to enhance our understanding of coronal jets and their role in solar and heliospheric dynamics.

To train and evaluate our random forest classifier, we divided the dataset into training and test sets with an 80-20 split. This approach ensures that the model is trained on a substantial portion of the data while preserving a separate set for unbiased performance evaluation.

Before training the model, we conducted a hyper-parameter tuning process leveraging a TPE. TPE is a sequential model-based optimization method [Bergstra et al., 2011] which leverages probability density functions to guide the search towards more promising regions of the hyperparameter space. This allows TPE to efficiently explore and exploit the search space, often leading to faster convergence to optimal solutions.

We employed such a method through the optimization framework Optuna[2] [Akiba et al., 2019]. The optimal settings are evaluated by means of k-fold cross-validation [Kohavi, 1995]. Once the model is optimized and trained, we evaluate its performance on the test set.

To evaluate classifier performance, we use multiple standard metrics, additionally we report confusion matrix. The confusion matrix, shown in Table 4.2, includes TP, TN, FP, and FN, which provide detailed insights into the model's predictions. Using multiple evaluation metrics offers a broader and more comprehensive understanding of the model's performance. This multi-metric approach helps identify strengths and weaknesses that a single metric might overlook. Evaluation scores are stored in table 4.3.

The accuracy score is 0.76, but it measures the ratio of correctly predicted instances to the total instances, and it can be misleading in unbalanced datasets where one class significantly outnumbers the other.

The balanced accuracy, with a score of 0.73, addresses this by evaluating the accuracy of each class individually and then averaging the results.

The model achieved an ROC-AUC score of 0.81. This metric represents the area under the receiver operating characteristic (ROC-AUC) curve, which plots the true positive rate (recall) against the false positive rate for various thresholds. This suggests that the model performs reasonably well across both classes. The high specificity indicates that the model effectively identifies negative cases, minimizing false positives. However, the recall and precision values reveal that there is room for improvement in correctly identifying positive cases, as it misses some positives and incorrectly labels some negatives as positives.

[2]Optuna documentation: https://optuna.readthedocs.io/en/stable/index.html

Figure 4.15 displays the distribution of correct and incorrect predictions across different thresholds. The green bar represents correct predictions (true positives and true negatives), and the red bar represents incorrect predictions (false positives and false negatives).

The plot shows that as the threshold increases, the confidence of the model in correctly predicting jets increases.

In this study, we aim to leverage machine learning to expand our sample of coronal jets. Hence, once the model is trained and validated, we apply it to classify the unlabeled data. The MM dataset, consisting of 876644 total structures, is extensive and diverse compared to the SAJIA dataset. Our primary goal is to obtain new samples of coronal jets, hence we adjust the prediction threshold from 0.5 to 0.95 to ensure the model outputs positive results only for instances with higher confidence. Such an approach led to the identification of 3452 new jet candidates. To further validate these results, we performed a manual inspection of each candidate by analyzing SDO/AIA 304 Å images in GIF format, captured near the date of the jet eruptions, to verify their authenticity.

Figure 4.17 presents an exemplary case where the new jet candidate is confirmed as a true coronal jet. Now, for the sake of clarity, let us take a closer look at the new coronal jets identified using the MM approach.

Figure 4.16 shows the distribution of the new detection in terms of intensity, time, latitude and area.

The density distributions of MM jets, SAJIA jets, and SAJIA Non-jets reveal several significant patterns. MM jets exhibit the highest densities at lower intensities and smaller areas, indicating that these jets are predominantly low-intensity, small-scale structures. Furthermore, MM jets are clustered primarily during the early stages of solar cycle 24, and are concentrated at high latitudes.

In contrast, while SAJIA jets and SAJIA Non-jets show a more gradual decline in density with increasing intensity and area, their distributions are more widespread across latitudes and over the entire time period examined. Importantly, the most populated areas for SAJIA jets coincide with the regions where MM jets are clustered, suggesting that MM jets tend to cluster in these highly populated areas of the training set.

To identify new jets, we utilized a machine-learning model with a decision threshold set at 0.95. This high threshold compels the model to select MM jets that are predominantly clustered in the most densely populated regions of the training set, which consists of SAJIA instances.

However, it is crucial to consider that these distributions may be influenced by the SDO intensity degradation effect [Ahmadzadeh et al., 2019, Barnes et al., 2020, Zwaard et al., 2021]. Over time, the degradation in intensity could impact the detection and classification of coronal jets. This degradation may lead to an overrepresentation of detections in certain regions and periods, potentially skewing the observed distributions. Therefore, while the data suggest notable trends, these biases should be accounted for when interpreting the results. After conducting the visual inspection of the candidate jets, we identified 3268 true jets and 184 false positives.

TABLE 4.3: Evaluation Metrics for random forest Classifier

Metric	**Score**
Accuracy	0.76
Balanced Accuracy	0.73
ROC-AUC	0.81
Precision	0.66
Recall	0.65
F1 Score	0.65
Specificity	0.82

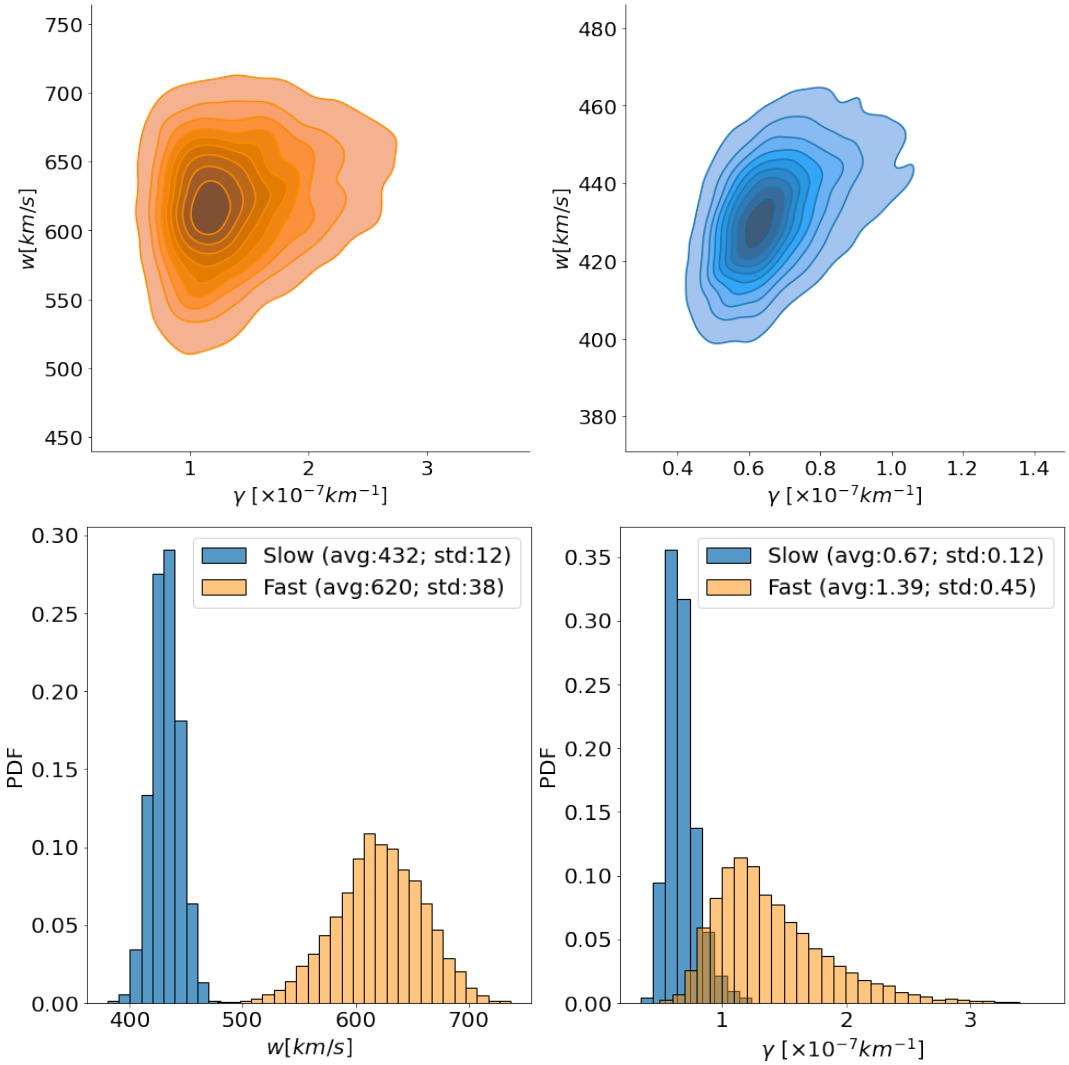


FIGURE 4.9: Posterior PDF obtained from the MCMC approach. (Upper left) Joint distribution of DBM parameters (γ, w) for the fast solar wind case. (Upper right) Joint distribution of DBM parameters (γ, w) for the slow solar wind case. Marginal PDF of γ (lower right) and w (lower left) for both fast and slow solar wind cases. The legend displays the average (avg) and standard deviation (std) values.

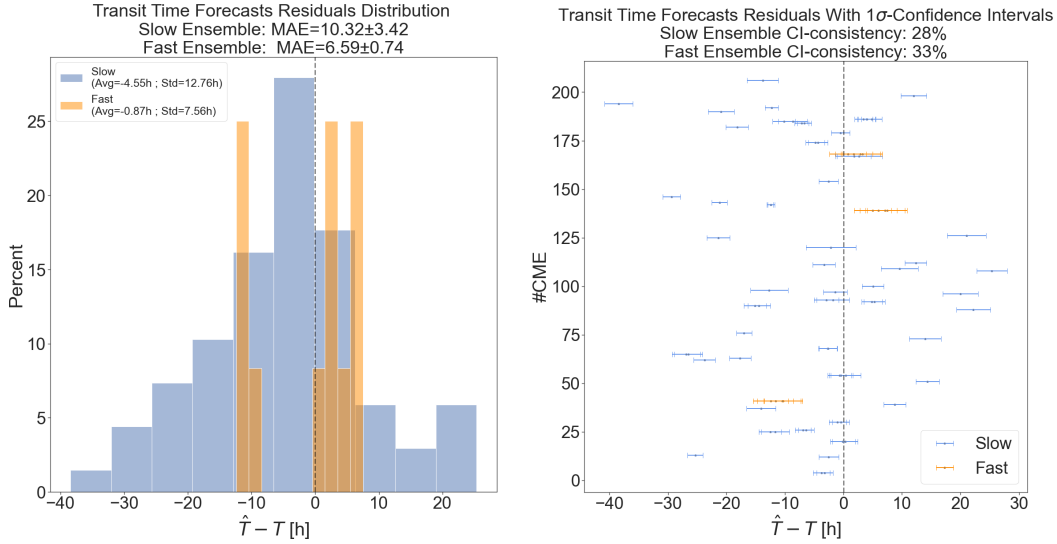


FIGURE 4.10: The transit time forecasting results using P-DBM with the ensemble approach. (Left) Histogram of the residuals $(\bar{T} - T)$, where $\bar{T}$ is the predicted transit time and T is the true transit time, providing an overview of the forecast error distribution. The legend indicates the mean and standard deviation of the residuals from four test folds. The mean value represents the average bias of the predictions, while the standard deviation reflects the variability of the errors. (Right) Scatter plot of the residuals $(\bar{T} - T)$ for each test CME, with associated error bars derived from P-DBM. The vertical axis corresponds to the CME number in the dataset, with each point representing an individual CME.

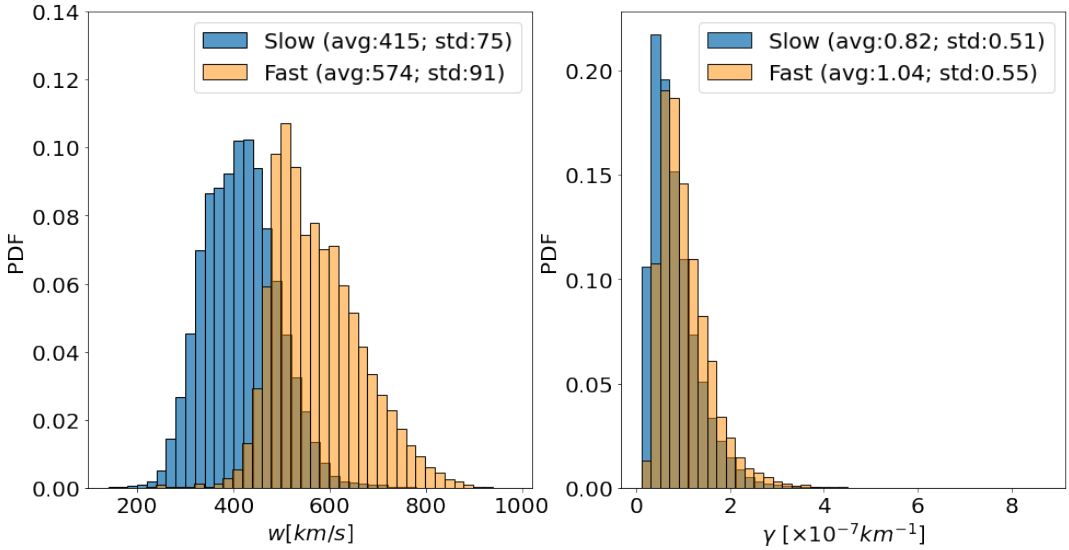


FIGURE 4.11: Histograms of marginal DBM Parameter PDF for the slow (Blue) and fast ensemble (Orange); obtained via individual approach.

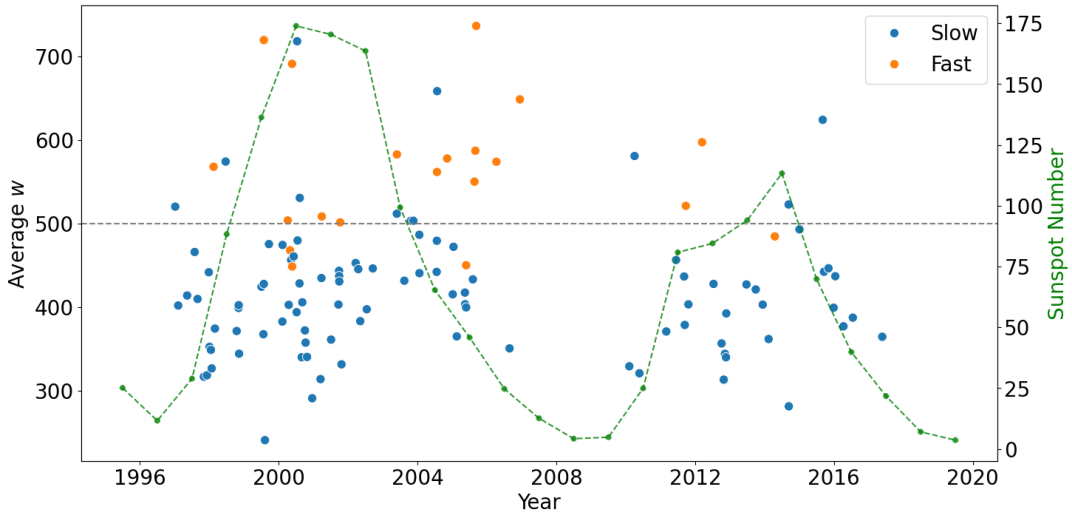


FIGURE 4.12: Scatter plot depicting the average solar wind speed (w) values of the PDFs obtained through the individual approach. CMEs labelled as slow and fast by Mugatwala et al. [2024] are shown as blue and orange dots, respectively. The second x-axis shows the line plot of the annually averaged Sunspot number (in green).

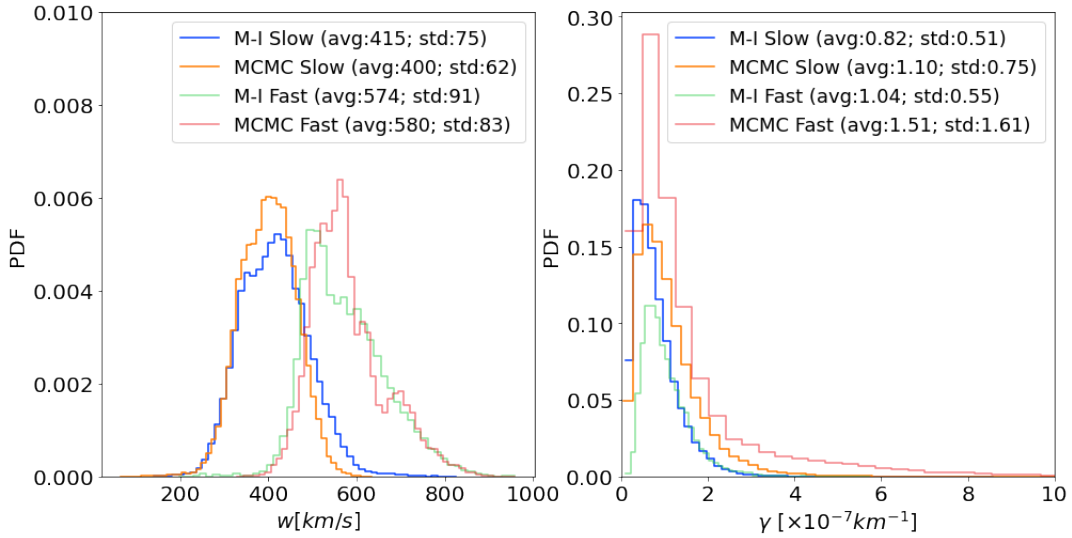


FIGURE 4.13: Histograms depicting the PDFs of marginal DBM parameters for the MCMC slow ensemble (MCMC Slow) and the MCMC fast ensemble (MCMC Fast) obtained via the individual approach. For comparison, the PDF of the ensembles from Mugatwala et al. [2024] (M-I Slow and M-I Fast) are also shown.

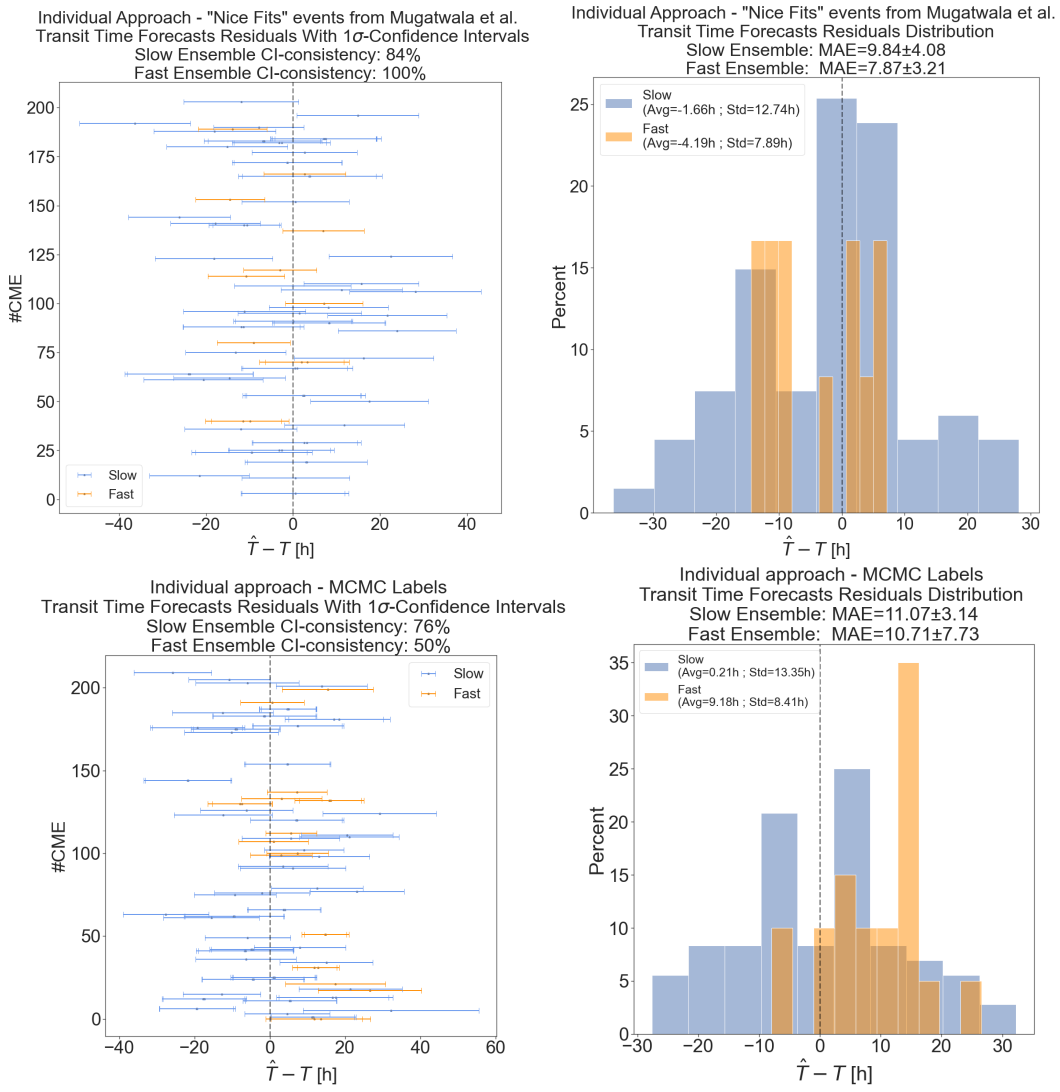


FIGURE 4.14: The transit time forecasting results with P-DBM obtained via individual approach. (right) Scatter-plot of the residuals $(\bar{T} - T)$ for all the test CMEs. (left) Histogram of the residuals $(\bar{T} - T)$ ($\bar{T}$ is the predicted transit time and T is the true transit time).

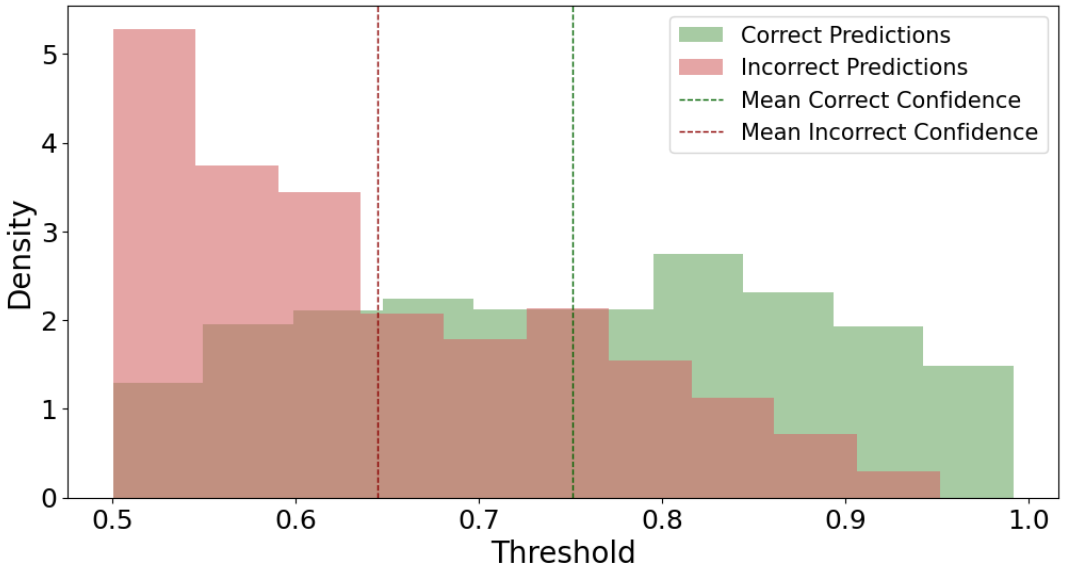


FIGURE 4.15: Confidence in correct vs incorrect predictions. Distribution of correct (green) and incorrect (red) predictions across different thresholds. The x-axis represents the thresholds ranging from 0.5 to 1.0, and the y-axis indicates the count of predictions.

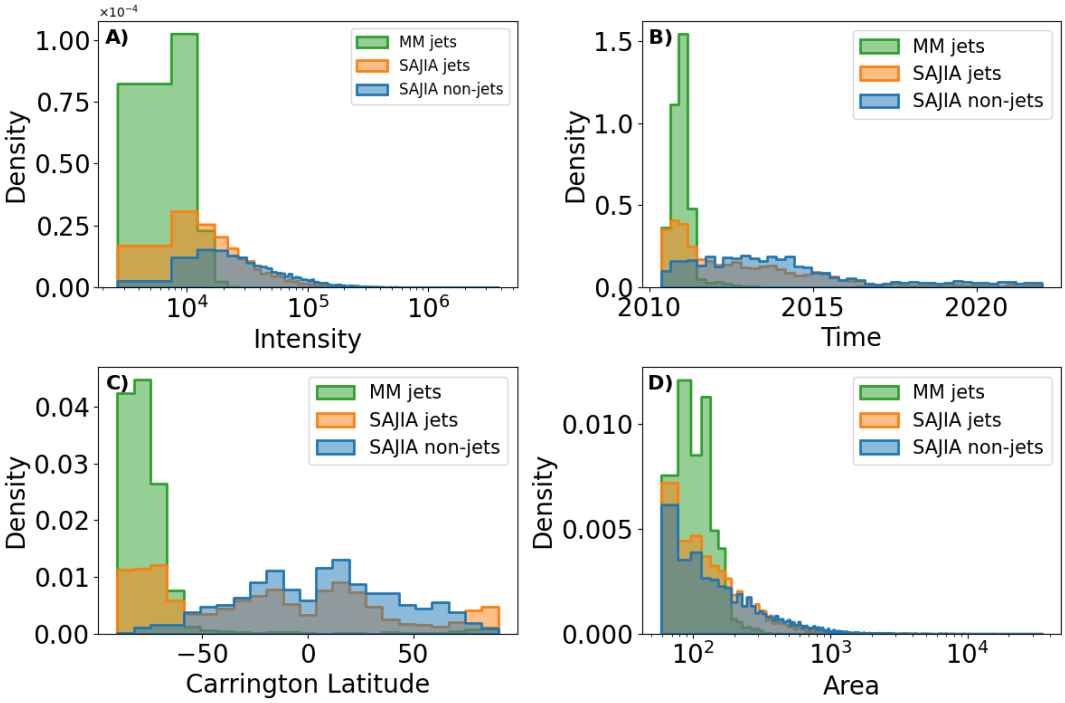


FIGURE 4.16: Figure shows the density distributions of MM jets, SAJIA jets, and SAJIA non-jets across four different features: Intensity (A), Time (B), Carrington Latitude (C), and Area (D). Each subplot shows the comparative density for each class, with MM jets indicated in green, SAJIA jets in orange, and SAJA non-jets in blue.

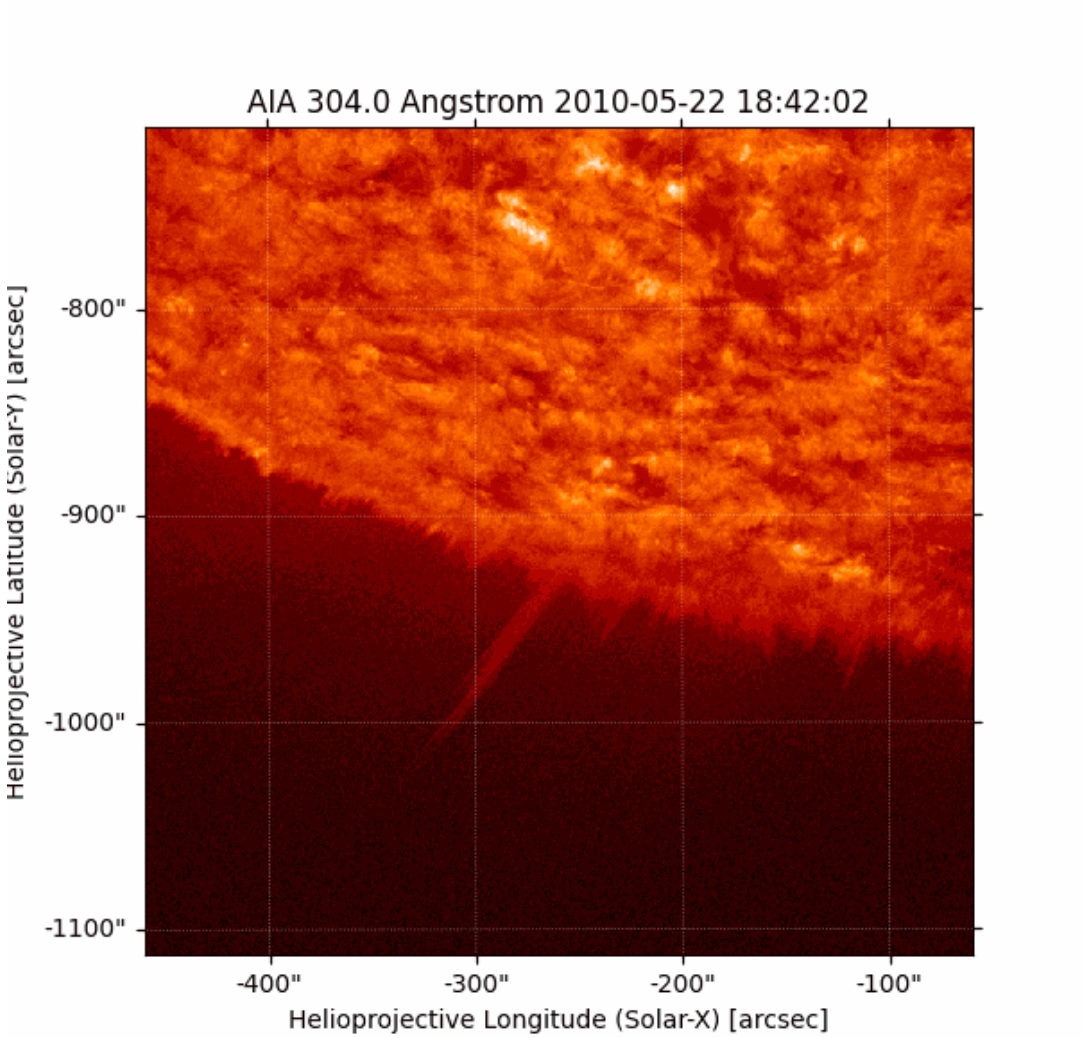


FIGURE 4.17: Confirmed true jet observed on May 22, 2010. The jet is visible as a bright, elongated structure extending from the solar surface into the upper atmosphere. The image is presented in helioprojective coordinates, with the x-axis representing helioprojective longitude (Solar-X) and the y-axis representing helioprojective latitude (Solar-Y), both in arcseconds.

Chapter 5

Discussion and conclusions

This thesis explored three distinct yet interrelated projects that apply ML, Bayesian inference, and mathematical morphology to improve understanding and forecasting of various space-weather phenomena. Although each project targeted a different aspect of solar and heliospheric physics—from CMEs to coronal jets—the unifying thread is the application of data-driven methods to address the limited accuracy, data availability, and complexity of current space-weather models.

In the following sections, we summarize each project's key contributions, discuss their limitations, and propose future research directions. We then synthesize overarching lessons learned and comment on the broader implications for space-weather forecasting.

Machine Learning for Earth-impacting CME prediction

Machine Learning continues to show strong potential in space weather research, especially in forecasting when and if CMEs will arrive at Earth. In our work, we delved deeper into the CAT-PUMA concept, examining how ML algorithms could leverage CME observations to improve Earth-impact predictions. Despite promising results in some cases, several challenges became apparent, underscoring the complexity of the problem and the limitations of current data.

One of the most pressing hurdles relates to the quantity and quality of the data used to train ML models. The relatively small number of documented Earth-Impacting CMEs makes it difficult to characterize the problem space comprehensively, which in turn limits the model's ability to generalize.

The relatively low number of well-characterised, Earth-impacting CMEs — typically only a few hundred — constrains the training of more sophisticated machine learning models and reduces the statistical reliability of predictive performance metrics. Ideally, robust model development would benefit from datasets comprising several thousand CME events with consistent labels, physical parameter annotations, and reliable arrival-time information. Whether such dataset sizes can be achieved through current or near-term missions remains uncertain. Earth-directed CMEs are intrinsically rare, and while ongoing missions such as Solar Orbiter and Parker Solar Probe provide valuable new insights, their contribution to significantly enlarging the pool of Earth-relevant training data may remain limited. In this context, complementary strategies such as harmonising historical event catalogues

across missions and exploring physics-informed data augmentation could offer more immediate and scalable paths toward improving dataset utility and model robustness.

Our regression analyses illustrated that ML performance can appear sufficiently high under certain training–test splits; however, cross-validation often revealed issues with overfitting and high variability in predictive accuracy. In particular, depending solely on BSV for model selection can inadvertently lead to *cherry-picked* data subsets, creating an unrealistic sense of the model's reliability.

Classification tasks also suffered from data-related constraints. Although we observed that some models could identify the dominant (non-Earth-impact) class with few mistakes, the FAR for Earth-impacting CMEs remained stubbornly above 70%. This phenomenon aligns with the findings of Fu et al. [2021] and other authors who reported similarly high false alarm rates, even when employing sophisticated models. Consistent with Vourlidas et al. [2019], our investigation indicates that high FAR values are not exclusive to ML-based approaches but also appear in MHD-driven models; however, the sparse data for Earth-impacting events exacerbates the problem in ML-focused research.

From a feature-engineering perspective, we focused primarily on CME speed, mass, and angular width as key descriptors, yet these variables only scratch the surface of CME interplanetary transport. Although CAT-PUMA incorporates elements intended to encode the state of the solar wind, our results—together with SHAP analyses—show that these features were either too coarse or too approximate to markedly improve the classification. This limitation is critical because the interplay between CMEs and the ambient solar wind strongly influences travel times and the likelihood of Earth impact. For instance, our assumption of a six-hour averaged SW speed at Lagrangian-1 (L1) may be insufficient to capture the dynamic and spatially evolving properties of the interplanetary medium.

Looking ahead, *enriching the feature space* stands out as a promising route for boosting forecasting skill. One strategy could be to integrate derived parameters, such as those in CAT-PUMA, with new features automatically extracted from white-light or EUV images by deep learning frameworks (Wang et al. [2019]; Fu et al. [2021]). Deep neural networks, in particular, can reduce the need for hand-crafted inputs by learning from raw images, though any such approach must carefully balance accuracy with the operational necessity of timely forecasts. Incorporating advanced geometric or physical descriptors of CMEs—for example, more precise information about their direction of propagation or three-dimensional structure—could further refine model predictions. Naturally, such expansions in input complexity may increase processing times and demand higher computational resources, an important practical consideration if we aim for near-real-time forecasting.

In summary, our exploration of ML within the CAT-PUMA framework reaffirms the strong potential of data-driven approaches in space weather, while drawing attention to the significant obstacles that must still be addressed. Data sparsity, imperfect measurements, and limited feature representations

consistently hinder model performance and forecasting reliability. Nevertheless, by deepening our understanding of the interplanetary environment and leveraging the continuous improvements in ML methodologies, we see room for substantial advances. In the broader context of this thesis, these findings resonate with the challenges noted in both the Bayesian and jet-detection projects, highlighting the indispensable roles of high-quality data, robust inference techniques, and carefully engineered feature spaces in pushing the boundaries of space-weather forecasting.

Bayesian approach to CME transit time forecasting

In this segment of the work, we explored how probabilistic methods could improve CME transit-time predictions, focusing on the P-DBM framework. We evaluated forecasting performance via a cross-validation scheme, where the P-DBM used the PDF of DBM parameters to generate transit-time predictions. Two complementary strategies were tested: (i) a group or ensemble approach, and (ii) an individual, per-event approach.

Using an MCMC algorithm, we derived posterior distributions for the DBM parameters by collectively fitting multiple CMEs. Only parameter sets satisfying the DBM equations *for all CMEs in the ensemble* were retained, effectively capturing their shared behavior under certain heliospheric conditions. In particular, we investigated how the drag parameter (γ) and solar wind speed (w) varied between slow and fast solar wind environments. Although these ensemble-derived PDFs offered decent performance in terms of mean absolute error (MAE), their probabilistic reliability was weaker, indicating that applying one-size-fits-all parameter constraints can mask the nuances of individual events.

By contrast, the individual approach the initial velocity v_0 as a free parameter, using weakly informative priors to maintain flexibility. We generated a specific PDF for each CME, later aggregating results into slow and fast solar wind categories. Despite showing MAE values comparable to the ensemble method, individual predictions carried larger error bars, reflecting greater uncertainty—but also more accurately capturing the variability among distinct CMEs.

Overall, we found that CMEs in slow solar wind tended to cluster around lower w values (roughly $w < 500$ km/s), while fast-wind events aligned with $w > 500$ km/s, consistent with several previous studies [e.g., Napoletano et al., 2018a, Mugatwala et al., 2024] except for the fast case in Napoletano et al. [2022], where the PDF of w averages 490 km/s.

However, the parameter γ proved more challenging to interpret, showing potential correlations with w in the ensemble approach that were less pronounced on an event-by-event basis. We suspect these relationships may partly stem from mathematical constraints embedded in the ensemble fitting.

Table 5.2 presents the results of CME transit time forecasting from this study, alongside results from other studies utilizing the DBM framework, including P-DBM and Drag based Ensemble model (DBEM). We also include results from machine learning models for broader comparison. Comparing

TABLE 5.1: The table presents the statistical moments (mean and standard deviation) of the distributions for the DBM parameters w and γ obtained in this study, along with a comparative analysis of similar findings from prior research.

Study	CME Ensemble	$\bar{w}$ $[km/s]$	σ_w $[km/s]$	$\bar{\gamma}$ $[\times 10^{-7} km^{-1}]$	σ_γ $[\times 10^{-7} km^{-1}]$
Napoletano et al. [2018a]	Slow	400	66	PDF for all CMEs	
	Fast	600	76	0.83	1.21
Napoletano et al. [2022]	Slow	370	80	PDF for all CMEs	
	Fast	490	100	0.96	3.62
Mugatwala et al. [2024]	Slow	371	89	0.86	0.80
	Fast	579	68	1.26	0.80
This work	Slow	432	12	0.67	0.12
(ensemble approach)	Fast	620	38	1.39	0.45
This work	M-I Slow	415	75	0.82	0.61
(individual approach)	M-I Fast	574	91	1.04	0.55
	MCMC Slow	400	62	1.10	0.75
	MCMC Fast	580	83	1.51	1.61

across studies is challenging due to the use of different datasets, criteria for model evaluation, and sample sizes. To provide context, we include additional information on the models, validation techniques, and test set sizes. Notably, the results for both the ensemble and individual approaches (M-I slow and fast) are based on the same training and test sets and evaluation methods, making them directly comparable.

TABLE 5.2: The table summarizes the mean MAE results achieved in this study for CME transit time forecasting and compares them with results from previous studies.

Study	Model	Validation method	Test size	MAE [h]
Napoletano et al. [2018a]	P-DBM	Hold-out	14	9.1
		Hold-out	100	16.8
Dumbović et al. [2018]	DBEM	Hold-out	25	14.3
Paouris et al. [2021]	DBEM	Hold-out	16	14.31 ± 2.18
Napoletano et al. [2022]	P-DBM	Hold-out	100	16.3
This work	P-DBM	4-fold CV	Slow - 17 [×4]	10.3 ± 3.4
(ensemble approach)			Fast - 3 [×4]	6.6 ± 0.7
This work	P-DBM	4-fold CV	M-I Slow - 17 [×4]	9.8 ± 4.1
(individual approach)			M-I Fast - 3 [×4]	7.9 ± 3.2
			MCMC Slow - 18 [×4]	11.1 ± 3.1
			MCMC Fast - 5 [×4]	10.7 ± 7.7
This work	CAT-PUMA Framework	5-fold CV	42 [×5]	>10
		Best hold-out (BSV)	42	7.6
Liu et al. [2018]	Support Vector Machines	Best hold-out (BSV)	37	5.9
Wang et al. [2019]	Convolutional Neural Network	10-fold CV	22 [×10]	12.4
Alobaid et al. [2022]	CMETNet	9-fold CV (adapted)	∼20	9.75
Guastavino et al. [2023]	Physics-driven NN	100 randomized splits	∼20	9.64

As with any data-driven technique, the quality and scope of the underlying datasets are critical. Recurring uncertainties in CME/ICME identification and measurement approximations can significantly affect posterior estimates, particularly for the sparser fast-wind events. Additionally, our binary classification of CMEs into only slow or fast may oversimplify the range of heliospheric states. This limitation was especially noticeable when the ensemble method struggled to capture the smaller sample of fast-wind CMEs, suggesting the need for more granular approaches or larger datasets.

Another key consideration is how strictly the ensemble approach forces parameter sets to satisfy the DBM equations for every CME in the group. While this produces a cohesive collective PDF, certain events that do not fit the model's assumptions can skew the results. On the other hand, the individual approach sacrifices some overall coherence to capture event-specific details, resulting in broader (though arguably more honest) PDFs.

Our findings point to several avenues for further research. Testing the algorithm on CMEs observed during different phases of the solar cycle could illuminate how γ and w shift over time and in varied heliospheric conditions. We also noted that the algorithm occasionally associates higher w with ascending or descending solar cycle periods, hinting at a link between solar-wind speed and cyclic changes in solar activity.

In addition, adapting more advanced MCMC strategies - for instance, the ensemble samplers proposed by Goodman and Weare [2010] - could improve the efficiency and accuracy of parameter estimation, yielding more robust PDF estimates. These techniques might offer a better balance between capturing inter-event variability and providing reliable ensemble constraints.

Despite its limitations, the DBM remains a widely used and computationally efficient tool for CME forecasting. By layering Bayesian methodologies on top of the basic drag-based framework, we gain both a clearer view of the uncertainties and a path toward refining transit-time predictions. Ongoing work to expand and improve the DBM characterization, coupled with smarter Bayesian sampling techniques, has the potential to strengthen our probabilistic forecasts of CME behavior and arrival times. In the broader context of this thesis, these insights align with the overarching theme of leveraging rigorous, data-driven approaches to tackle fundamental questions of space-weather forecasting and operational reliability.

Coronal jet detection through Machine learning and mathematical morphology

In this work, we focused on augmenting our coronal jet dataset by applying a random forest model and incorporating outputs from both the SAJIA algorithm and a MM approach. Our primary objective was to enrich the existing catalog of coronal jets and thereby deepen our understanding of these features. By fusing SAJIA detections with MM-based geometric information, we constructed a more comprehensive dataset that captures key structural aspects of jet phenomena.

It is worth noting that the implementation of the MM algorithm in this study was specifically optimised for the detection of off-limb coronal jets, which are more easily distinguishable due to enhanced contrast and reduced background interference. However, the method is not inherently limited to off-limb events. With suitable adaptations—such as tailored background subtraction, contrast enhancement, and refined structuring elements—the MM framework could, in principle, be extended to detect on-disk jets as well. Such an extension would be particularly valuable for constructing a more complete inventory of jet activity across the solar surface. Future work may

explore this direction, enabling more comprehensive statistical studies and broader applicability of the jet detection pipeline.

The inclusion of MM features added vital shape and size descriptors, ultimately boosting the classification model's ability to differentiate genuine jets from other solar phenomena. After training and validating the random forest on a suitably large sample, we used the classifier to label previously unclassified data with a carefully chosen threshold aimed at minimizing false positives. This process yielded a total of 3452 new jet candidates, which we then verified through a manual inspection of corresponding Graphics Interchange Formats (GIFs) thereby confirming 3268 true jets and identifying 184 false alarms. These results highlight not only the effectiveness of integrating ML methods with analytical techniques but also the power of combining automated detection with visual validation.

Beyond expanding the coronal jet dataset, our analysis suggests that harnessing machine learning in tandem with classic image-processing strategies can significantly amplify the rate at which new solar features are discovered and cataloged. This enlargement of jet statistics is particularly relevant when investigating broader solar phenomena, such as active longitudes [Chidambara Aiyar, 1932, Plyusnina, 2010, Zhang et al., 2008, Gyenge et al., 2017], which remain a topic of ongoing debate. More extensive jet observations may also play an important role in understanding the solar dynamo and, by extension, improving space-weather forecasting models.

Despite these successes, we emphasize again the critical role of data quality in sustaining reliable ML performance. As demonstrated in the other projects (e.g., CAT-PUMA and P-DBM-based approaches), inaccurate or incomplete data can yield misleading results, particularly in tasks requiring high-confidence classifications. Ensuring sufficiently broad, well-labeled, and representative datasets stands as a key priority for future progress in coronal jet detection. Looking ahead, our pipeline could benefit further from deep learning architectures that reduce dependence on manual feature engineering and speed up the identification of newly emerging jets.

Taken together with the outcomes of the previous projects, these findings show that advanced data-driven methods can be successfully adapted to a range of solar-physics challenges, from forecasting CMEs to systematically cataloging eruptive jets. By continuously refining both the data and the algorithms, we can enhance our capability to monitor and interpret solar phenomena, thereby contributing to more robust, data-rich space-weather prediction frameworks.

Synopsis of key contributions and outlook

This thesis investigates how ML and Bayesian inference can improve our ability to model the arrival times and effects of CMEs, which are central drivers of space-weather phenomena. By combining supervised learning techniques (such as SVMs, decision trees, and ensemble methods) with probabilistic drag-based models enriched by MCMC, the research aims to enhance our predictive understanding of CME propagation and mitigate the

risk these events pose to Earth's technological infrastructure. An additional focus on coronal jets broadens the scope of solar phenomena under study, illustrating how data-driven approaches can inform various eruptive events on the Sun.

Several noteworthy findings emerge from this effort. First, by training ML algorithms on CME observations, the work demonstrates a more robust approach for predicting CME arrival times. Techniques such as SHAP values offer detailed insights into which features (for example, CME speed, angular width, or solar wind indicators) most strongly influence the model's predictions, thereby increasing both interpretability and user confidence. Second, Bayesian inference, implemented through a revised Metropolis-Hastings algorithm, refines our understanding of CME behavior by explicitly quantifying uncertainties and providing more reliable estimates of transit times. Finally, the thesis applies a random forest classifier, combined with mathematical morphology, to augment coronal jet catalogs, illuminating how ML can systematically identify rare or subtle phenomena in solar datasets. Together, these advances address both data scarcity and uncertainty management, thereby promoting more transparent and reliable space-weather forecasting models.

Solar activity, and CMEs in particular, can severely disrupt satellite communication, power grids, and other vital technologies. The methods developed here provide new tools to tackle such challenges. By creating more accurate and interpretable CME predictions, risk mitigation strategies can be employed more efficiently, reducing the potential for system failures or service interruptions. While traditional physics-based methods, such as MHD simulations, remain valuable, the ML and Bayesian frameworks introduced in this thesis afford greater flexibility and can integrate diverse data sources, from coronagraph imagery to in-situ solar wind measurements at L1. Moreover, augmenting jet datasets through ML contributes to a broader understanding of solar eruptive phenomena and complements ongoing research into topics like solar dynamo processes and active longitudes. Taken together, these contributions form a more holistic, data-driven view of space-weather forecasting than typically available in the literature.

Despite the promise of ML and deep learning methods, several factors decisively influence their effectiveness in space-weather research. Data quality and diversity remain essential, given that sparse or noisy measurements can lead to biased or unstable predictions. This reinforces the importance of comprehensive, well-labeled, and balanced datasets, which can also benefit from synthetic data generation if managed with attention to physical realism. Equally critical are the evaluation methodologies employed: metrics like precision, recall, F1-score, AUC, and MAE should align with the specific modeling goals, whether classification, regression, or probabilistic inference. Furthermore, model interpretability is crucial for scientific scrutiny; if decision-making processes are opaque, results may be met with skepticism, particularly in operational environments. Architectural and algorithmic considerations also matter. Deep learning architectures and transfer-learning strategies can extract hierarchical features from complex solar data, but they must be

carefully tailored to the available observations and computational resources, especially when near-real-time forecasts are needed. Bayesian methods, similarly, benefit from more specialized MCMC variants, such as ensemble samplers, that can handle high-dimensional solar data without sacrificing too much computational efficiency.

Building on the foundational work presented in this thesis, several promising pathways for future research in space weather forecasting emerge. A key opportunity lies in expanding the use of multi-wavelength observations, including X-ray, EUV, and white-light coronagraph data, to better capture the intricate details of CME formation and propagation. These rich datasets, coupled with advanced ML paradigms—particularly deep learning models designed for spatiotemporal data—offer significant potential for enhancing predictive accuracy.

In particular, specific features extracted from EUV and white-light observations—such as coronal dimmings, EUV wave fronts, and pre-eruptive structures like cavities and sigmoidal loops—have been shown to correlate with key CME parameters and may serve as valuable inputs for predictive models [Aschwanden, 2010, Dissauer et al., 2019, Harrison et al., 2003, Zhang et al., 2007]. Coronal dimmings, for instance, are strongly associated with CME mass and speed, while EUV waves often indicate the onset and lateral extent of eruptions [Thompson et al., 1998, Muhr et al., 2011]. Sigmoidal loop morphologies and coronal cavities are also linked with eruptive potential, as they typically reflect highly sheared or flux-rope-dominated magnetic topologies [Sarkar et al., 2019, Savcheva et al., 2015]. The emergence of missions like Parker Solar Probe and Solar Orbiter further expands the horizon of feature accessibility. By providing multi-perspective, high-resolution observations of the low corona and inner heliosphere, instruments such as SoloHI, Metis, and EUI can help overcome projection effects, capture early eruption signatures, and refine estimates of CME dynamics [Rouillard et al., 2016, Andretta et al., 2021, Howard et al., 2020]. Incorporating such physically grounded features into data-driven pipelines may support more robust, generalisable, and operationally relevant forecasting frameworks.

Techniques such as transfer learning and domain adaptation can address the challenge of data scarcity by leveraging knowledge from related tasks or larger, more accessible datasets.

Transfer learning, for example, allows models pre-trained on large-scale datasets, such as ImageNet [Deng et al., 2009], to be fine-tuned for specific space weather tasks. By adapting generalised feature extraction capabilities, this approach reduces the need for extensive task-specific data and computational resources [Pan and Yang, 2010, Tan et al., 2018, Zhuang et al., 2019]. Recent studies, such as Upendran et al. [2020], demonstrate the effectiveness of transfer learning in predicting solar wind speeds at L1. Their model utilises EUV images from NASA's SDO, specifically leveraging the 193 Å wavelength for its sensitivity to coronal holes and the 211 Å wavelength for its focus on active regions. Such applications underscore the potential of transfer learning to enable accurate predictions in data-constrained environments.

Another promising direction involves leveraging deep learning to integrate

diverse data sources and capture the complex, multi-scale dynamics of space weather. For example, deep neural networks have shown great potential in predicting solar wind speed at Earth. Models that process SDO/EUV image sequences use CNNs for spatial feature extraction, combined with attention-based modules to capture critical temporal and spatial patterns. These models predict key solar wind properties such as speed, density, and, potentially, magnetic field components, contributing to a deeper understanding of solar wind dynamics.

Recent advancements, such as those presented by Brown et al. [2022], highlight the effectiveness of attention-based architectures like Vision Transformer (ViT) for solar wind speed forecasting. By processing sequences of image patches, these models dynamically assess spatial relationships and temporal evolution, identifying patterns such as the development and movement of coronal holes. Advanced architectures like Transformer in Transformer (TNT) and Swin Transformer [Han et al., 2021, Liu et al., 2021] further enhance the capacity to handle time-dependent data by building hierarchical feature maps and processing patches within patches. These methods excel at capturing solar activity's complex temporal evolution, particularly during the declining solar cycle phase when coronal holes dominate. Studies have consistently shown that attention-based models outperform traditional convolutional approaches in these scenarios, reinforcing their value for future applications.

Moreover, generative AI represents a promising frontier in space weather research, offering innovative solutions to longstanding challenges such as data scarcity and class imbalance in solar flare forecasting. The work of Ramunno et al. [2024] exemplifies this potential by introducing Denoising Diffusion Probabilistic Models (DDPMs) to generate synthetic solar images with controlled solar flare intensities. By addressing the rarity of high-energy events like M- and X-class flares, this approach not only balances datasets but also enhances the training and performance of machine learning models for solar activity classification and prediction. Their findings demonstrate that generative AI can outperform traditional data augmentation techniques, ensuring more accurate and diverse datasets while maintaining physical relevance. This highlights how generative models can advance both predictive capabilities and fundamental understanding of solar phenomena, paving the way for future research applications in heliophysics and beyond.

Another area ripe for exploration is uncertainty quantification. Bayesian inference and sampling strategies can enhance the interpretability and reliability of predictions, providing probabilistic outputs that are more actionable for real-world applications. Incorporating these approaches into deep learning models could help bridge the gap between scientific advancements and operational forecasting.

While the machine learning models developed in this thesis — including the CAT-PUMA-inspired regressors — offer speed and adaptability, they do not currently account for the observational uncertainties associated with input CME parameters, such as speed, angular width, or mass, which are often subject to measurement error and inter-catalogue variability. In contrast, the probabilistic framework implemented via MCMC sampling in

the P-DBM model was specifically designed to propagate uncertainty in the forecast outputs, such as time of arrival and arrival speed, by sampling from empirically derived input distributions. This approach enables the generation of both point estimates and credible intervals, thereby enhancing forecast reliability and transparency. Looking ahead, the development of machine learning models that are intrinsically uncertainty-aware represents a promising research direction. Probabilistic deep learning methods — including Bayesian neural networks, Monte Carlo dropout [Gal and Ghahramani, 2016], and ensemble-based approaches [Lakshminarayanan et al., 2017] — provide mechanisms to model both epistemic uncertainty (due to limited training data) and aleatoric uncertainty (due to noise in observations). These techniques are gaining traction in scientific domains where understanding model confidence is critical [Abdar et al., 2021]. Additionally, hybrid frameworks that integrate physical constraints — such as drag-based propagation models — with deep neural networks are emerging as powerful tools for improving data efficiency while maintaining physical plausibility [Reichstein et al., 2024, Karniadakis et al., 2021]. As space weather datasets continue to grow in volume and richness, embedding uncertainty quantification into forecasting frameworks will be vital for improving predictive skill and supporting the transition to reliable, operational deployment.

Operational integration will require efficient and maintainable pipelines for data ingestion, model training, and prediction delivery. Technologies such as cloud computing and containerization can streamline these processes, ensuring that deep learning models are scalable and ready for real-time applications. For instance, embedding these techniques into operational forecasting systems could revolutionise our ability to provide timely and accurate predictions, mitigating the disruptive impacts of space weather on critical infrastructure.

In this context, it is worth noting that the machine learning models explored in this thesis—such as random forests and CAT-PUMA-style regressors—exhibit low computational demands: training typically completes within minutes, and inference times are well below one second per event on a standard desktop machine. This computational efficiency makes them well suited for real-time deployment scenarios.

This thesis underscores the transformative potential of blending ML, Bayesian inference, and domain expertise in space-weather forecasting. By emphasizing model interpretability, robust validation, and thoughtful data handling, it provides a methodological blueprint for grappling with the complexities of CMEs and related solar eruptions. The findings set the stage for more accurate, trustworthy, and ultimately actionable predictions, strengthening our collective capability to anticipate and mitigate the risks of an active Sun. Future endeavors that embrace deeper neural networks, generative data augmentation, and real-time operational pipelines hold the promise of further refining both predictive power and scientific rigor, ensuring that society's critical infrastructure remains resilient against disruptive solar events.

Appendix A

Bayesian Method

A.1 Fundamental Concepts of Bayesian Theory

In this section, we will highlight some fundamental concepts of Bayesian theory [Regis, 2015, Ivezić et al., 2019]. Probability is commonly defined as the quantification of the degree of randomness or uncertainty associated with an event. In general, probability is a measure of the likelihood or chance of various events occurring, which is evident from our everyday usage of the term "probability". The concept of probability is intrinsically linked to the notion of uncertainty. Probability can be viewed as a function of an event that produces a numerical value representing the likelihood or chance of that event happening. There are multiple ways to define and calculate such a probability function [Kolmogorov and Bharucha-Reid, 2018].

Historically, the differing interpretations of the concept of probability stem from its dual significance — epistemic and empirical. The epistemic conception of probability considers the uncertainty linked to the concept of probability, which arises from the limited and imperfect nature of human knowledge and understanding. This conception acknowledges that our ability to predict and model probabilistic events is constrained by gaps and biases in our knowledge, as well as the inherent complexities and unpredictabilities of the phenomena we study. In contrast, the empirical conception views uncertainty as an intrinsic and irreducible characteristic of phenomena themselves, existing independently of human knowledge. This perspective holds that even with complete information, certain events and processes possess inherent randomness that cannot be fully eliminated or predicted [Jaynes, 2003].

Bayesian statistics originates from an epistemic perspective, where uncertainty stems from incomplete knowledge of a fundamentally deterministic system. This perspective acknowledges underlying causal mechanisms but recognizes our ability to model them is limited by the information available. The Bayesian approach provides a formal framework for updating beliefs about the system as new information becomes available, allowing for iterative refinement of knowledge and reduced uncertainty over time [Eddy, 2004].

Probability can, in fact, be defined in three different ways. The *classical definition* states that the probability of an event is the ratio between the number of favorable cases and the number of possible cases, assuming all events are equally probable [Marquis de Laplace, 1902]. The *frequentist definition*

describes probability as the limit of the frequency of an event when the number of observations, N, tends to infinity [Mises, 2013]. Lastly, the *Bayesian definition* interprets probability as a measure of the degree of credibility of a proposition [Jaynes, 2003]. Each approach has inherent limitations due to the assumptions they rest upon.

The classical conception of probability, which assumes discrete and finite events, faces challenges when applied to continuous variables. A key limitation is the presumption of perfect uniformity, wherein all possible outcomes are known *a priori* and equally likely, introducing circularity in the definition [Haack and Duica, 1993]. This assumption renders the classical framework ill-suited for more complex systems where equal likelihood cannot be consistently presumed.

Conversely, the frequentist interpretation defines probability in terms of the relative frequency of an event's occurrence in repeated trials. This definition is linked to the law of large numbers, which asserts that the experimental frequency will approach the true probability as the number of trials, N, tends to infinity [Kolmogorov and Bharucha-Reid, 2018]:

$$\lim_{N \to +\infty} \mathcal{P}\left[\left(\frac{N_E}{N} - \mathcal{P}(E)\right) < \epsilon\right] = 1, \tag{A.1}$$

where ϵ is an arbitrarily small positive number. This formulation implies that the probability of the frequency of the event deviating from its true value by more than ϵ becomes increasingly small as N grows. However, the frequentist approach assumes the experiment is repeatable under identical conditions, which may not always be feasible [Feller, 1991].

The Bayesian approach differs from classical and frequentist methods by incorporating subjective beliefs about the likelihood of an event occurring. This is achieved by assuming an *a priori* distribution that represents the degree of credibility assigned to a hypothesis before any data is observed [Bernardo and Smith, 2009]. Bayesian inference updates this prior distribution using observed data to compute a posterior distribution, quantifying the updated belief about the hypothesis in light of evidence.

A central tenet of Bayesian theory is Bayes' theorem, which provides a systematic way to update a-priori probabilities with new empirical evidence. By adjusting initial beliefs based on observed frequencies, Bayes' theorem enables calculation of an a-posteriori probability, a revised assessment of the hypothesis's credibility after accounting for data. This interplay between subjective beliefs and objective observations is a hallmark of the Bayesian approach, enabling probability assessments even in scenarios lacking directly relevant frequency data. Bayesian probability theory is grounded in axiomatic principles governing coherent assignment and manipulation of probabilities [Bayes and Price, 1763, Hanke et al., 2014].

Referring to $\mathcal{S}$ as the *space of events* (i.e. the set of all the possible results of an experiment) and considering an event A (i.e. a subset of $\mathcal{S}$, $A \subset \mathcal{S}$), the probability $\mathcal{P}$ associated to A is a real number such that:

Axiom 1. *For any event A, $\mathcal{P}(A) \geq 0$ (a negative probability has no meaning).*

Axiom 2. *If $\mathcal{S}$ is the sample space for a given experiment, $\mathcal{P}(\mathcal{S}) = 1$ (probabilities are normalized so that the maximum value is unity).*

Axiom 3. *If $A \cap B = \oslash$, then $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B)$. More generally, For an infinite number of mutually exclusive sets A_i, $i = 1, 2, 3...$ ($A_i \cap Aj = \oslash$ for all $i \neq j$),*

$$\mathcal{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathcal{P}(A_i).$$

In particular, the second part of Axiom 3 is to be taken from the following corollary

Corollary A.1.0.1. Consider M sets A_1, A_2, ..., A_M which are mutually exclusive, $A_i \cap Aj = \oslash$ for all $i \neq j$,

$$\mathcal{P}(\bigcup_{i=1}^{M} A_i) = \sum_{i=1}^{M} \mathcal{P}(A_i)$$

Proof. The proof is given using mathematical induction. it is noted that by Axiom 3, the statement applies for M = 2, and hence it must be true for M = 3. Since it is true for M = 3, it must also be true for M = 4, and so on. In this way, we can prove that Corollary A.1.0.1 is true for any finite M. □

From these axioms, the entire theory of probability can be developed.

A.1.1 Bayes' Theorem

This subsection is devoted to the introduction of one of the fundamental results of Bayesian theory, namely Bayes' theorem. The main properties of probability theory that characterise the Bayesian approach are the concepts of joint probability and conditional probability.

Definition A.1.1 (Joint Probability). Given two events, A and B, the *joint probability* of events A and B (typically denoted by $\mathcal{P}(A, B) \equiv \mathcal{P}(A \cap B)$) is defined as the probability that the events occur simultaneously.

Definition A.1.2 (Conditional Probability). For two events A and B , the probability of A conditioned on knowing that B has occurred is

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A, B)}{\mathcal{P}(B)}$$

The notion of event A given event B does not mean that event B has occurred (e.g. is certain); instead, it is the probability of event A occurring after or in the presence of event B for a given trial.

We now show some results with the aim of arriving at the formulation of Bayes' theorem.

Theorem A.1.1. For any events A and B such that $\mathcal{P}(B) \neq 0$,

$$\mathcal{P}(A, B) = \frac{\mathcal{P}(B|A)\mathcal{P}(A)}{\mathcal{P}(B)}$$

Proof. From Def. A.1.2:

$$\mathcal{P}(A, B) = \mathcal{P}(A|B)\mathcal{P}(B) = \mathcal{P}(B|A)\mathcal{P}(A).$$

Theorem A.1.1 follows directly by dividing the preceding equations by $\mathcal{P}(B)$. □

Theorem A.1.2 (Theorem of Total Probability). Let $B_1, B_2, ..., B_n$ be a set of mutually exclusive and exhaustive events. That is, $B_i \cup B_j = \oslash$ for all $i \neq j$ and

$$\bigcup_{i=1}^{n} B_i = \mathcal{S} \Rightarrow \sum_{i=1}^{n} \mathcal{P}(B_i) = 1$$

Then

$$\mathcal{P}(A) = \sum_{i=1}^{n} \mathcal{P}(A|B_i)\mathcal{P}(B_i)$$

Proof. the event A can be written as

$$A = \{A \cup B_1\} \cap \{A \cup B_2\} \cup ... \cup \{A \cap B_n\}$$

Also, since the B_i are all mutually exclusive, then the $\{A \cup B_i\}$ are also mutually exclusive so that

$$\begin{aligned} \mathcal{P}(A) &= \sum_{i=1}^{n} \mathcal{P}(A, B_i) \qquad \text{(by Corollary A.1.0.1),} \\ &= \sum_{i=1}^{n} \mathcal{P}(A|B_i)\mathcal{P}(B_i) \qquad \text{(by Theorem A.1.1).} \end{aligned}$$

□

Bayes' theorem will be introduced below, this theorem basically gives a relation between conditional probabilities. The validity of this theorem derives essentially from the definition of conditional probability and the combination the results of Theorems A.1.1 and A.1.2 represents the foundation of the Bayesian method [Bayes, 1958, Gelman et al., 1995, Bishop and Nasrabadi, 2006, Bernardo and Smith, 2009].

Theorem A.1.3 (Bayes's Theorem). Let $B_1, B_2, ..., B_n$ be a set of mutually exclusive and exhaustive events. Then,

$$\mathcal{P}(B_i|A) = \frac{\mathcal{P}(A|B_i)\mathcal{P}(B_i)}{\sum_{i=1}^{n} \mathcal{P}(A|B_i)\mathcal{P}(B_i)}.$$

$\mathcal{P}(B_i)$ is often referred to as the *a-priori* probability of event B_i, while $\mathcal{P}(B_i|A)$ is known as the *a-posteriori* probability of event B_i given A [Bayes, 1958, Gelman et al., 1995].

The reason why this theorem is considered so fundamental to the construction of Bayesian theory is that it allows the use of the machinery of probability theory to describe the uncertainty in model parameters, often denoted by the symbol $\boldsymbol{\theta}$ (or $\boldsymbol{w}$), or indeed in the choice of the model itself [Bernardo and Smith, 2009]. In more detail, it allows capturing the assumptions about $\boldsymbol{\theta}$, before observing the data, in the form of a prior probability distribution $\mathcal{P}(\boldsymbol{\theta})$, then the effect of the observed data $\mathcal{D} = \{d_{t_1}, ..., d_{t_N}\}$ is expressed through the conditional probability $\mathcal{P}(\mathcal{D}|\boldsymbol{\theta})$.

Bayes' theorem, which takes the form:

$$\mathcal{P}(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta})}{\mathcal{P}(\mathcal{D})}, \tag{A.2}$$

then allows us to evaluate the uncertainty in $\boldsymbol{\theta}$ after $\mathcal{D}$ has been observed in the form of the posterior probability $\mathcal{P}(\boldsymbol{\theta}|\mathcal{D})$. The quantity $\mathcal{P}(\mathcal{D}|\boldsymbol{\theta})$ on the right-hand side of Bayes' theorem is evaluated for the observed dataset $\mathcal{D}$ and can be viewed as a function of the parameter vector $\boldsymbol{\theta}$, in which case it is called the *Likelihood function* [Berger, 1985, Bishop and Nasrabadi, 2006].

Bayesian priors

$\mathcal{P}(\boldsymbol{\theta})$ is called the *prior distribution* because it does not take into account any information regarding experimental data ($\mathcal{D}$). It therefore represents a sort of bias placed before measurements are even made [Bayes, 1958, Bernardo and Smith, 2009]. The prior incorporates all other knowledge that might exist, but is not used when computing the likelihood and therefore can include the knowledge extracted from prior measurements of the same type as the data at hand.

For example, we may know from older work that the value m_A of the mass of an elementary particle, with a Gaussian uncertainty parametrized by σ_A, but we wish to utilize a new measuring apparatus or method. Hence, m_A and σ_A may represent a convenient summary of the posterior PDF from older work that is now used as a prior for the new measurements [Gelman et al., 1995]. Therefore, the terms prior and posterior do not have an absolute meaning. Such priors that incorporate information based on other measurements (or other sources of meaningful information) are called *informative priors* [Bolstad and Curran, 2016]. When no other information, except for the data we are analyzing, is available, one possibility is to assign priors by formal rules. Sometimes these priors are called *uninformative priors* but, despite the misleading name, these priors can incorporate weak but objective information such as "the model parameter describing variance cannot be negative" [Jaynes, 2003]. Note that even the most uninformative priors still affect the estimates, and the results are not generally equivalent to other inference approaches.

Although uninformative priors do not contain specific information, they can be assigned according to several general principles. These principles are

formulated under the belief that the same prior information should result in the assignment of the same priors. A few examples are given below:

- *Principle of indifference*: A set of basic, mutually exclusive possibilities needs to be assigned equal probabilities [Keynes, 2013]. An example could be the case of a fair six-sided die, where each of the outcomes has a prior probability of 1/6.
- *Principle of consistency*: The prior for a location parameter should not change with translations of the coordinate system and yields a flat prior. Similarly, the prior for a scale parameter should not depend on the choice of units [Bernardo, 1979].

In addition, when we have additional weak prior information about some parameter, such as a low-order statistic, we can use the *principle of maximum entropy* to construct priors consistent with that information [Jaynes, 1957].

A.1.2 Likelihood Function

Given an independent and identically distributed (iid) sample $\mathcal{D}n = (d_1, ..., d_n)$ from a density $f\boldsymbol{\theta}$, with an unknown set of parameters $\boldsymbol{\theta} = \theta_1, \theta_2, ..., \theta_p$, where $\theta_i \in \Theta$ (an example could be the mean μ and variance σ of a Gaussian distribution), the associated *likelihood function* is

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} l_{\boldsymbol{\theta}}(x_i). \tag{A.3}$$

This quantity is a fundamental entity for the analysis of the information provided about the parameter $\boldsymbol{\theta}$ by the sample $\mathcal{D}_n$, and Bayesian analysis relies on this function to draw inference on $\boldsymbol{\theta}$ [Berger, 1985, Gelman et al., 1995].

This quantity can be viewed as a function of the parameter vector $\boldsymbol{\theta}$ and it expresses how probable the observed data set is for different settings of the parameter vector $\boldsymbol{\theta}$ [Robert et al., 2007]. Note that the likelihood is not a probability distribution over $\boldsymbol{\theta}$, and its integral with respect to $\boldsymbol{\theta}$ does not (necessarily) equal one.

The major input of the Bayesian perspective is that it modifies the likelihood, which is a simple function of $\boldsymbol{\theta}$, into a posterior distribution on the parameter $\boldsymbol{\theta}$. In this sense, the likelihood is transformed into an a-posteriori distribution, dependent on the parameter $\boldsymbol{\theta}$ defined by

$$\mathcal{P}(\boldsymbol{\theta}|\mathcal{D}_n) = \frac{\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}_n)\mathcal{P}(\boldsymbol{\theta})}{\int \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}_n)\mathcal{P}(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{A.4}$$

The above likelihood offers the dual interpretation of the probability density of $\mathcal{D}_n$ conditional on the parameter $\boldsymbol{\theta}$, with the additional indication that the observations in $\mathcal{D}_n$ are independent given $\boldsymbol{\theta}$ [Gelman et al., 1995]. The numerator in (A.4) is therefore the joint density on the pair ($\mathcal{D}_n$, $\boldsymbol{\theta}$), and the Bayes theorem provides the conditional (or posterior) distribution of the

parameter $\boldsymbol{\theta}$ given the sample $\mathcal{D}_n$. The denominator is called the marginal (likelihood) $\mathcal{M}(\mathcal{D}_n)$ [Berger, 1985].

Bayesian Inference

This subsection provides a concise overview of the key conceptual steps in the Bayesian inference process. Statistical inference refers to the procedure of drawing conclusions from model estimations. The Bayesian approach extends traditional statistical inference by allowing probability statements not only about data but also about model parameters and the models themselves [Berger, 1985, Jaynes, 2003]. In this framework, inferences are made by producing PDFs, treating model parameters as random variables rather than fixed quantities [Gelman et al., 1995].

At the core of Bayesian inference is Bayes' theorem, which formalizes the process of continually updating knowledge about a phenomenon [Bayes, 1958]. Initially, prior information is encoded in the prior distribution, which represents beliefs about the parameters before observing any data. Once the data $\mathcal{D}$ is observed, the likelihood function is multiplied by the prior to produce the posterior distribution, representing the updated knowledge about the parameters after considering the data [Bishop and Nasrabadi, 2006]. Importantly, when new data are collected, the posterior from the previous analysis can serve as the prior for the subsequent analysis, facilitating an iterative refinement of knowledge.

The Bayesian inference process involves the formulation of the likelihood function, $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$, which represents the probability of the observed data given the model parameters. The next step is the selection of the prior, $\mathcal{P}(\boldsymbol{\theta})$, which incorporates any relevant prior knowledge or beliefs about the parameters that are not directly related to the data. Finally, the posterior distribution, $\mathcal{P}(\boldsymbol{\theta}|\mathcal{D})$, is calculated using Bayes' theorem by combining the likelihood and the prior [Bayes, 1958, Bayes and Price, 1763, Gelman et al., 1995].

Once the posterior distribution is computed, the next task is to estimate the model parameters $\boldsymbol{\theta}$ that maximize the posterior probability. This is commonly achieved using the Maximum a Posteriori (MAP) estimate, given by:

$$\boldsymbol{\theta}MAP = \arg\max \boldsymbol{\theta}\mathcal{P}(\mathcal{D}|\boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta}), \tag{A.5}$$

which yields the parameter values that are most likely given the observed data and prior information [Bishop and Nasrabadi, 2006].

Beyond obtaining a point estimate, Bayesian inference also focuses on quantifying the uncertainty of these estimates. This is done by constructing credible intervals, which provide a range of values within which the true parameter value is likely to lie. In the one-dimensional case, the $(1-\alpha)$-level credible region is determined by finding values a and b such that:

$$\int_{-\infty}^{a} \mathcal{P}(\theta|\mathcal{D})d\theta = \int_{b}^{\infty} \mathcal{P}(\theta|\mathcal{D})d\theta = \frac{\alpha}{2}. \tag{A.6}$$

Thus, the probability that the true value of θ lies within the interval (a, b) is $1 - \alpha$, and this interval is referred to as the $(1 - \alpha)$ posterior interval [Gelman et al., 1995]. The process of integrating over the probability distribution to focus on specific parameters is known as marginalization, and the resulting distribution is called the marginal posterior PDF.

In cases where the model involves multiple parameters, the joint posterior distribution provides the a posteriori probability for all parameters $\boldsymbol{\theta}$. To analyze individual parameters, marginalization is used to reduce the joint posterior to a distribution for a specific parameter θ_i:

$$\mathcal{P}(\theta_i|\mathcal{D}) = \int \mathcal{P}(\boldsymbol{\theta}|\mathcal{D}) d\theta_1 d\theta_2 \dots d\theta_m, \tag{A.7}$$

where the integration is performed over all other parameters except θ_i [Robert et al., 2007]. Marginalization is also valuable for understanding covariances between parameters, as it allows for integration over so-called nuisance parameters, which are not of primary interest. This process facilitates the focus on parameters of scientific importance while accounting for the uncertainty in the remaining parameters.

In summary, Bayesian inference provides a powerful framework for parameter estimation and uncertainty quantification, where both the data and prior information are incorporated in a cohesive probabilistic manner, and conclusions are continually updated as new information becomes available [Bayes, 1958, Gelman et al., 1995].

Bibliography

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.05.008. URL https://www.sciencedirect.com/science/article/pii/S1566253521001081.

Azim Ahmadzadeh, Dustin J. Kempton, and Rafal A. Angryk. A curated image parameter data set from the solar dynamics observatory mission. *The Astrophysical Journal Supplement Series*, 243(1):18, jul 2019. doi: 10.3847/1538-4365/ab253a. URL https://dx.doi.org/10.3847/1538-4365/ab253a.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

Doha Al-Feadh and Wathiq Al-Ramdhan. Large geomagnetic storms drives by solar wind in solar cycle 24. *IOP Publishing*, 1234(1):012004–012004, 07 2019. doi: 10.1088/1742-6596/1234/1/012004. URL https://doi.org/10.1088/1742-6596/1234/1/012004.

Hannes Alfvén. Existence of electromagnetic-hydrodynamic waves. *Nature*, 150(3805):405–406, 1942.

Khalid A. Alobaid, Yasser Abduallah, Jason T. L. Wang, Haimin Wang, Haodi Jiang, Yan Xu, Vasyl Yurchyshyn, Hongyang Zhang, Huseyin Cavus, and Ju Jing. Predicting CME arrival time through data integration and ensemble learning. *Frontiers in Astronomy and Space Sciences*, 9:1013345, October 2022. doi: 10.3389/fspas.2022.1013345.

B. L. Alterman, Y. J. Rivera, S. T. Lepri, and R. M. Raines. On the transition from Slow to Fast Wind as Observed in Composition Observations. *arXiv e-prints*, art. arXiv:2411.18984, November 2024. doi: 10.48550/arXiv.2411.18984.

T. Amari, J. F. Luciani, J. J. Aly, Z. Mikic, and J. Linker. Coronal mass ejection: Initiation, magnetic helicity, and flux ropes. II. turbulent diffusion–driven

evolution. *The Astrophysical Journal*, 595(2):1231–1250, October 2003. doi: 10.1086/377444. URL https://doi.org/10.1086/377444.

V. Andretta, A. Bemporad, Y. De Leo, G. Jerse, F. Landini, M. Mierla, G. Naletto, M. Romoli, C. Sasso, A. Slemer, D. Spadaro, R. Susino, D. C. Talpeanu, D. Telloni, L. Teriaca, M. Uslenghi, E. Antonucci, F. Auchère, D. Berghmans, A. Berlicki, G. Capobianco, G. E. Capuano, C. Casini, M. Casti, P. Chioetto, V. Da Deppo, M. Fabi, S. Fineschi, F. Frassati, F. Frassetto, S. Giordano, C. Grimani, P. Heinzel, A. Liberatore, E. Magli, G. Massone, M. Messerotti, D. Moses, G. Nicolini, M. Pancrazzi, M. G. Pelizzo, P. Romano, U. Schühle, M. Stangalini, Th. Straus, C. A. Volpicelli, L. Zangrilli, P. Zuppella, L. Abbo, R. Aznar Cuadrado, R. Bruno, A. Ciaravella, R. D'Amicis, P. Lamy, A. Lanzafame, A. M. Malvezzi, P. Nicolosi, G. Nisticò, H. Peter, C. Plainaki, L. Poletto, F. Reale, S. K. Solanki, L. Strachan, G. Tondello, K. Tsinganos, M. Velli, R. Ventura, J. C. Vial, J. Woch, and G. Zimbardo. The first coronal mass ejection observed in both visible-light and UV H I Ly-α channels of the Metis coronagraph on board Solar Orbiter. *Astronomy and Astrophysics*, 656:L14, December 2021. doi: 10.1051/0004-6361/202142407.

Andreea Anghel, Nikolaos Papandreou, T. A. Parnell, Alessandro De Palma, and Haralampos Pozidis. Benchmarking and optimization of gradient boosting decision tree algorithms, 01 2018. URL https://arxiv.org/abs/1809.04559.

Yuichiro Anzai. *Pattern recognition and machine learning*. Elsevier, 2012.

Markus J. Aschwanden. *Physics of the Solar Corona. An Introduction with Problems and Solutions (2nd edition)*. Springer Science & Business Media, 2005.

Markus J. Aschwanden. Image Processing Techniques and Feature Recognition in Solar Physics. *Solar Physics*, 262(2):235–275, April 2010. doi: 10.1007/s11207-009-9474-y.

Mariette Awad and Rahul Khanna. Support vector machines for classification, 01 2015. URL https://doi.org/10.1007/978-1-4302-5990-9_3.

Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Elsevier BV*, 59:44–58, 07 2020. doi: 10.1016/j.inffus.2020.01.005. URL https://doi.org/10.1016/j.inffus.2020.01.005.

Badruddin, F. Mustajab, and M. Derouich. Geomagnetic response of interplanetary coronal mass ejections in the earth's magnetosphere. *Elsevier BV*, 154:1–4, 05 2018. doi: 10.1016/j.pss.2018.01.012. URL https://doi.org/10.1016/j.pss.2018.01.012.

Salome R Bagashvili, Bidzina M Shergelashvili, Darejan R Japaridze, Vasil Kukhianidze, Stefaan Poedts, Teimuraz V Zaqarashvili, Maxim L Khodachenko, and Patrick De Causmaecker. Evidence for precursors of the

coronal hole jets in solar bright points. *The Astrophysical Journal Letters*, 855 (2):L21, 2018. doi: 10.3847/2041-8213/aab08b.

John Bahcall, Marc Pinsonneault, and Sarbani Basu. Solar models: Current epoch and time dependences, neutrinos, and helioseismological properties. *The Astrophysical Journal*, 555, 11 2000. doi: 10.1086/321493.

D. N. Baker. Effects of the sun on the earth's environment. *Elsevier BV*, 62(17-18):1669–1681, 11 2000. doi: 10.1016/s1364-6826(00)00119-x. URL https://doi.org/10.1016/s1364-6826(00)00119-x.

D. N. Baker, X. Li, A. Pulkkinen, Chigomezyo M. Ngwira, M. L. Mays, A. B. Galvin, and K. D. C. Simunac. A major solar eruptive event in july 2012: Defining extreme space weather scenarios. *American Geophysical Union*, 11 (10):585–591, 10 2013. doi: 10.1002/swe.20097. URL https://doi.org/10.1002/swe.20097.

S. D. Bale, S. T. Badman, J. W. Bonnell, T. A. Bowen, D. Burgess, A. W. Case, C. A. Cattell, B. D. G. Chandran, C. C. Chaston, C. H. K. Chen, J. F. Drake, T. Dudok de Wit, J. P. Eastwood, R. E. Ergun, W. M. Farrell, C. Fong, K. Goetz, M. Goldstein, K. A. Goodrich, P. R. Harvey, T. S. Horbury, G. G. Howes, J. C. Kasper, P. J. Kellogg, J. A. Klimchuk, K. E. Korreck, V. V. Krasnoselskikh, S. Krucker, R. Laker, D. E. Larson, R. J. MacDowall, M. Maksimovic, D. M. Malaspina, J. Martinez-Oliveros, D. J. McComas, N. Meyer-Vernet, M. Moncuquet, F. S. Mozer, T. D. Phan, M. Pulupa, N. E. Raouafi, C. Salem, D. Stansby, M. Stevens, A. Szabo, M. Velli, T. Woolley, and J. R. Wygant. Highly structured slow solar wind emerging from an equatorial coronal hole. *Nature*, 576(7786):237–242, December 2019. doi: 10.1038/s41586-019-1818-7.

Andre Balogh, H. Hudson, K. Petrovay, and Rudolf Von Steiger. Introduction to the solar activity cycle: Overview of causes and consequences. *Space Science Reviews*, 186:1–15, 12 2014. doi: 10.1007/s11214-014-0125-8.

T. Barata, S. Carvalho, I. Dorotovic, F. Pinheiro, A. Garcia, J. Fernandes, and A. M. Lourenco. Software tool for automatic detection of solar plages in the Coimbra observatory spectroheliograms. *Astronomy and Computing*, 2018. doi: 10.48550/ARXIV.1811.08389. URL https://arxiv.org/abs/1811.08389.

Will Barnes, Mark Cheung, Monica Bobra, Paul Boerner, Georgios Chintzoglou, Drew Leonard, Stuart Mumford, Nicholas Padmanabhan, Albert Shih, Nina Shirman, David Stansby, and Paul Wright. aiapy: A Python Package for Analyzing Solar EUV Image Data from AIA. *The Journal of Open Source Software*, 5(55):2801, November 2020. doi: 10.21105/joss.02801.

Mr. Bayes and Mr. Price. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Royal Society*, 53:370–418, 12 1763. doi: 10.1098/rstl.1763.0053. URL https://doi.org/10.1098/rstl.1763.0053.

Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3-4):296–315, 1958.

S. M. Bennett and R. Erdélyi. On the Statistics of Macrospicules. *Astrophysical Journal*, 808(2):135, August 2015. doi: 10.1088/0004-637X/808/2/135.

Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology, 10 2008. URL https://doi.org/10.1371/journal.pcbi.1000173.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer New York, NY, 2 edition, 1985. doi: 10.1007/978-1-4757-4286-2. URL https://doi.org/10.1007/978-1-4757-4286-2. Originally published with the title: Statistical Decision Theory. Springer Book Archive. Copyright Springer Science+Business Media New York 1985.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Jose M Bernardo. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2): 113–128, 1979.

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

Prantika Bhowmik and Dibyendu Nandy. Prediction of the strength and timing of sunspot cycle 25 reveal decadal-scale space environmental conditions. *Nature Communications*, 9:5209, December 2018. doi: 10.1038/s41467-018-07690-0.

L. Biermann. Physical Processes in Comet Tails and their Relation to Solar Activity. In P. Swings, editor, *Liege International Astrophysical Colloquia*, volume 4 of *Liege International Astrophysical Colloquia*, pages 251–262, January 1952.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

William M Bolstad and James M Curran. *Introduction to Bayesian statistics*. John Wiley & Sons, 2016.

S. Bourgeois, S. Chierichini, Sz. Soós, R. Erdélyi, J. Liu, M. B. Korsós, R. Gafeira, and T. Barata. Long-term properties of coronal off-limb structures. *Astronomy and Astrophysics*, 693:A301, January 2025. doi: 10.1051/0004-6361/202451257.

Slava Bourgeois, Teresa Barata, R. Erdélyi, Ricardo Gafeira, and Orlando Oliveira. Sunspots identification through mathematical morphology. *Solar Physics*, 299, 01 2024. doi: 10.1007/s11207-023-02243-1.

S. Bravo and G. A. Stewart. Fast and Slow Wind from Solar Coronal Holes. *Astrophysical Journal*, 489(2):992–999, November 1997. doi: 10.1086/304789.

Stephen P. Brooks. Markov chain monte carlo method and its application. *Wiley*, 47(1):69–100, 03 1998. doi: 10.1111/1467-9884.00117. URL https://doi.org/10.1111/1467-9884.00117.

Edward J. E. Brown, Filip Svoboda, Nigel P. Meredith, Nicholas Lane, and Richard B. Horne. Attention-Based Machine Vision Models and Techniques for Solar Wind Speed Forecasting Using Solar EUV Images. *Space Weather*, 20(3):e2021SW002976, March 2022. doi: 10.1029/2021SW00297610.1002/essoar.10508581.1.

GE Brueckner, J-P Delaboudiniere, RA Howard, SE Paswaters, OC St. Cyr, R Schwenn, P Lamy, GM Simnett, B Thompson, and D Wang. Geomagnetic storms caused by coronal mass ejections (cmes): March 1996 through june 1997. *Geophysical Research Letters*, 25(15):3019–3022, 1998.

Natalia Buzulukova and Bruce Tsurutani. Space Weather: From Solar Origins to Risks and Hazards Evolving in Time. *Frontiers in Astronomy and Space Sciences*, 9:429, December 2022. doi: 10.3389/fspas.2022.1017103.

Enrico Camporeale, Simon Wing, and Jay Johnson. *Machine learning techniques for space weather*. Elsevier, 2018a.

Enrico Camporeale, Simon Wing, and Jay R. Johnson. *Copyright*. Elsevier, 2018b. ISBN 978-0-12-811788-0. doi: https://doi.org/10.1016/B978-0-12-811788-0.09994-7. URL https://www.sciencedirect.com/science/article/pii/B9780128117880099947.

Enrico Camporeale, Simon Wing, and Jay R. Johnson. Introduction. In Enrico Camporeale, Simon Wing, and Jay R. Johnson, editors, *Machine Learning Techniques for Space Weather*, pages xiii–xviii. Elsevier, 2018c. ISBN 978-0-12-811788-0. doi: https://doi.org/10.1016/B978-0-12-811788-0.09987-X. URL https://www.sciencedirect.com/science/article/pii/B978012811788009987X.

Richard C. Canfield, Kevin P. Reardon, K. D. Leka, K. Shibata, T. Yokoyama, and M. Shimojo. H alpha Surges and X-Ray Jets in AR 7260. *Astrophysical Journal*, 464:1016, June 1996. doi: 10.1086/177389.

Olivier Cappé and Christian P. Robert. Markov chain monte carlo: 10 years and still running! *Journal of the American Statistical Association*, 95(452):1282–1286, 12 2000. doi: 10.1080/01621459.2000.10474330. URL https://doi.org/10.1080/01621459.2000.10474330.

Peter J Cargill. On the aerodynamic drag force acting on interplanetary coronal mass ejections. *Solar Physics*, 221(1):135–149, 2004.

H. Carmichael. *A Process for Flares*, volume 50, page 451. Scientific and Technical Information Office, National Aeronautics and Space . . . , 1964.

R. C. Carrington. On Dr. Sœmmering's Observations of the Solar Spots, in the years 1826-1829. *Monthly Notices of the Royal Astronomical Society*, 20:71, January 1860. doi: 10.1093/mnras/20.3.71.

S. Carvalho, S. Gomes, T. Barata, A. Lourenço, and N. Peixinho. Comparison of automatic methods to detect sunspots in the Coimbra observatory spectroheliograms. *Astronomy and Computing*, 32:100385, 2020. ISSN 2213-1337. doi: https://doi.org/10.1016/j.ascom.2020.100385. URL https://www.sciencedirect.com/science/article/pii/S2213133720300391.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

M. Cécere, A. Costa, H. Cremades, and G. Stenborg. Recent insights on CME deflections at low heights. *Frontiers in Astronomy and Space Sciences*, 10: 1260432, September 2023. doi: 10.3389/fspas.2023.1260432.

Sumanjit Chakraborty, S. Ray, Dibyendu Sur, Abhirup Datta, and A. Paul. Effects of cme and cir induced geomagnetic storms on low-latitude ionization over indian longitudes in terms of neutral dynamics. *Elsevier BV*, 65(1):198–213, 01 2020. doi: 10.1016/j.asr.2019.09.047. URL https://doi.org/10.1016/j.asr.2019.09.047.

Paul Charbonneau. Dynamo Models of the Solar Cycle. *Living Reviews in Solar Physics*, 7(1):3, September 2010. doi: 10.12942/lrsp-2010-3.

Huadong Chen, Jun Zhang, Bart De Pontieu, Suli Ma, B. Kliem, and E. R. Priest. Coronal mini-jets in an activated solar tornado-like prominence. *IOP Publishing*, 899(1):19–19, 08 2020. doi: 10.3847/1538-4357/ab9cad. URL https://doi.org/10.3847/1538-4357/ab9cad.

PF Chen. Coronal mass ejections: models and their observational basis. *Living Reviews in Solar Physics*, 8(1):1–92, 2011.

P. R. Chidambara Aiyar. Two Longitudinal Zones of Apparent Inhibition of Sunspots on the Solar Disc. *Monthly Notices of the Royal Astronomical Society*, 93(2):150–151, 12 1932. ISSN 0035-8711. doi: 10.1093/mnras/93.2.150. URL https://doi.org/10.1093/mnras/93.2.150.

Simone Chierichini, Gregoire Francisco, Ronish Mugatwala, Raffaello Foldes, Enrico Camporeale, Giancarlo De Gasperis, Luca Giovannelli, Gianluca Napoletano, Dario Del Moro, and Robertus Erdelyi. A Bayesian approach to the drag-based modelling of ICMEs. *Journal of Space Weather and Space Climate*, 14:1, January 2024a. doi: 10.1051/swsc/2023032.

Simone Chierichini, Jiajia Liu, Marianna B. Korsós, Dario Del Moro, and Robertus Erdélyi. CME Arrival Modeling with Machine Learning. *Astrophysical Journal*, 963(2):121, March 2024b. doi: 10.3847/1538-4357/ad1cee.

L. P. Chitta, A. N. Zhukov, D. Berghmans, Hardi Peter, S. Parenti, Sudip Mandal, R. Aznar Cuadrado, U. Schühle, L. Teriaca, F. Auchère, Krzysztof Barczyński, E. Buchlin, L. K. Harra, E. Kraaikamp, David M. Long, L. Rodríguez, Conrad Schwanitz, Phil Smith, C. Verbeeck, and Daniel B. Seaton. Picoflare jets power the solar wind emerging from a coronal hole on the sun. *American Association for the Advancement of Science*, 381(6660):867–872, 08 2023. doi: 10.1126/science.ade5801. URL https://doi.org/10.1126/science.ade5801.

D. D. Clayton. Book-Review - Principles of Stellar Evolution and Nucleosynthesis. *Astronomy Express*, 1:81, September 1984.

Edward W Cliver, Carolus J Schrijver, Kazunari Shibata, and Ilya G Usoskin. Extreme solar events. *Living Reviews in Solar Physics*, 19(1):1–143, 2022.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995. doi: 10.1007/BF00994018.

Steven R. Cranmer. Coronal Holes. *Living Reviews in Solar Physics*, 6(1):3, September 2009. doi: 10.12942/lrsp-2009-3.

Steven R. Cranmer, Craig E. DeForest, and Sarah E. Gibson. Inward-propagating plasma parcels in the solar corona: Models with aerodynamic drag, ablation, and snowplow accretion. *The Astrophysical Journal*, 913(1):4, may 2021. doi: 10.3847/1538-4357/abf146. URL https://dx.doi.org/10.3847/1538-4357/abf146.

S. Dasso, C. H. Mandrini, P. Démoulin, and M. L. Luoni. A new model-independent method to compute magnetic helicity in magnetic clouds. *Astronomy and Astrophysics*, 455(1):349–359, August 2006. doi: 10.1051/0004-6361:20064806.

Jennifer L. Davidson and Gerhard X. Ritter. Theory of morphological neural networks. In Raymond Arrathoon, editor, *Digital Optical Computing II*, volume 1215 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 378–388, July 1990. doi: 10.1117/12.18085.

Bart De Pontieu, Robert Erdélyi, and Stewart P. James. Solar chromospheric spicules from the leakage of photospheric oscillations and flows. *Nature*, 430(6999):536–539, July 2004. doi: 10.1038/nature02749.

Dario Del Moro, Gianluca Napoletano, Roberta Forte, Luca Giovannelli, Ermanno Pietropaolo, and Francesco Berrilli. Forecasting the 2018 february 12th cme propagation with the p-dbm model: A fast warning procedure. *Annals of Geophysics*, 62(4):GM456–GM456, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

S. Derivaux, S. Lefevre, C. Wemmert, and J. Korczak. On machine learning in watershed segmentation. In *2007 IEEE Workshop on Machine Learning for Signal Processing*, pages 187–192, 2007. doi: 10.1109/MLSP.2007.4414304.

Sahel Dey, Piyali Chatterjee, Murthy O. V. S. N., Marianna B. Korsós, Jiajia Liu, Christopher J. Nelson, and Robertus Erdélyi. Polymeric jets throw light on the origin and nature of the forest of solar spicules. *Nature Physics*, 18(5): 595–600, March 2022. doi: 10.1038/s41567-022-01522-1.

Andrea Diercke, Robert Jarolim, Christoph Kuckein, Sergio J. González Manrique, Marco Ziener, Astrid M. Veronig, Carsten Denker, Werner Pötzi, Tatiana Podladchikova, and Alexei A. Pevtsov. A universal method for solar filament detection from h-alpha observations using semi-supervised deep learning, 2024.

Thomas G. Dietterich. Ensemble methods in machine learning, 01 2000. URL https://doi.org/10.1007/3-540-45014-9_1.

K. Dissauer, A. M. Veronig, M. Temmer, and T. Podladchikova. Statistics of Coronal Dimmings Associated with Coronal Mass Ejections. II. Relationship between Coronal Dimmings and Their Associated CMEs. *Astrophysical Journal*, 874(2):123, April 2019. doi: 10.3847/1538-4357/ab0962.

J. F. Drake, O. Agapitov, M. Swisdak, S. T. Badman, S. D. Bale, T. S. Horbury, J. C. Kasper, R. J. MacDowall, F. S. Mozer, T. D. Phan, M. Pulupa, A. Szabo, and M. Velli. Switchbacks as signatures of magnetic flux ropes generated by interchange reconnection in the corona. *Astronomy and Astrophysics*, 650: A2, June 2021. doi: 10.1051/0004-6361/202039432.

Mateja Dumbović, Jaša Čalogović, Bojan Vršnak, Manuela Temmer, M Leila Mays, Astrid Veronig, and Isabell Piantschitsch. The drag-based ensemble model (dbem) for coronal mass ejection propagation. *The Astrophysical Journal*, 854(2):180, 2018. doi: 10.3847/1538-4357/aaaa66.

Tibor Durgonics, A. Komjáthy, O. P. Verkhoglyadova, E. B. Shume, H. H. Benzon, A. J. Mannucci, Mark D. Butala, Per Høeg, and Richard B. Langley. Multiinstrument observations of a geomagnetic storm and its effects on the arctic ionosphere: A case study of the 19 february 2014 storm. *Wiley-Blackwell*, 52(1):146–165, 01 2017. doi: 10.1002/2016rs006106. URL https://doi.org/10.1002/2016rs006106.

Sean R. Eddy. What is bayesian statistics?, 09 2004. URL https://doi.org/10.1038/nbt0904-1177.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with

noise. In *kdd*, volume 96, pages 226–231, 1996. URL https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf.

William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons, 1991.

T. G. Forbes. A review on the genesis of coronal mass ejections, 10 2000. URL https://doi.org/10.1029/2000ja000005.

Gianni Franchi, Amin Fehri, and Angela Yao. Deep morphological networks. *Pattern Recognition*, 102:107246, June 2020. doi: 10.1016/j.patcog.2020.107246. URL https://hal.archives-ouvertes.fr/hal-02922299.

Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joëlle Pineau. An introduction to deep reinforcement learning. *Now Publishers*, 11(3-4):219–354, 01 2018. doi: 10.1561/2200000071. URL https://doi.org/10.1561/2200000071.

Emma Fry. The risks and impacts of space weather: Policy recommendations and initiatives. *Elsevier BV*, 28(3):180–184, 08 2012. doi: 10.1016/j.spacepol.2012.06.005. URL https://doi.org/10.1016/j.spacepol.2012.06.005.

Huiyuan Fu, Yuchao Zheng, Yudong Ye, Xueshang Feng, Chaoxu Liu, and Huadong Ma. Joint geoeffectiveness and arrival time prediction of cmes by a unified deep learning framework. *Remote Sensing*, 13(9):1738, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

Sergio González, Salvador García, Javier Del Ser, Lior Rokach, and Francisco Herrera. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Elsevier BV*, 64:205–237, 12 2020. doi: 10.1016/j.inffus.2020.07.007. URL https://doi.org/10.1016/j.inffus.2020.07.007.

Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5(1):65–80, 2010. doi: 10.2140/camcos.2010.5.65.

N. Gopalswamy. Extreme solar eruptions and their space weather consequences, 01 2018. URL https://doi.org/10.1016/b978-0-12-812700-1.00002-9.

N. Gopalswamy, Z. Mikić, D. Maia, David Alexander, H. Cremades, P. Kaufmann, Durgesh Tripathi, and Y.-M. Wang. The pre-cme sun. *Springer Science+Business Media*, 123(1-3):303–339, 10 2006. doi: 10.1007/s11214-006-9020-2. URL https://doi.org/10.1007/s11214-006-9020-2.

Nat Gopalswamy. Properties of Interplanetary Coronal Mass Ejections. *Space Science Reviews*, 124(1-4):145–168, June 2006. doi: 10.1007/s11214-006-9102-1.

Nat Gopalswamy, A Lara, RP Lepping, ML Kaiser, D Berdichevsky, and OC St. Cyr. Interplanetary acceleration of coronal mass ejections. *Geophysical research letters*, 27(2):145–148, 2000.

John T Gosling. The solar wind, 11 2006. URL https://www.sciencedirect.com/science/article/pii/B9780120885893500098.

LM Green, SA Matthews, L van Driel-Gesztelyi, LK Harra, and JL Culhane. Multi-wavelength observations of an x-class flare without a coronal mass ejection. *Solar Physics*, 205(2):325–339, 2002.

Lucie M. Green, Tibor Török, B. Vršnak, W. B. Manchester, and Astrid Veronig. The origin, early evolution and predictability of solar eruptions. *Springer Science+Business Media*, 214(1), 02 2018a. doi: 10.1007/s11214-017-0462-5. URL https://doi.org/10.1007/s11214-017-0462-5.

Lucie M. Green, Tibor Török, Bojan Vršnak, Ward Manchester, and Astrid Veronig. The origin, early evolution and predictability of solar eruptions. *Space Science Reviews*, 214(1), February 2018b. ISSN 1572-9672. doi: 10.1007/s11214-017-0462-5. URL http://dx.doi.org/10.1007/s11214-017-0462-5.

Sabrina Guastavino, Valentina Candiani, Alessandro Bemporad, Francesco Marchetti, Federico Benvenuto, Anna Maria Massone, Salvatore Mancuso, Roberto Susino, Daniele Telloni, Silvano Fineschi, and Michele Piana. Physics-driven Machine Learning for the Prediction of Coronal Mass Ejections' Travel Times. *Astrophysical Journal*, 954(2):151, September 2023. doi: 10.3847/1538-4357/ace62d.

N. Gyenge, T. Singh, T. S. Kiss, A. K. Srivastava, and R. Erdélyi. Active Longitude and Coronal Mass Ejection Occurrences. *Astrophysical Journal*, 838(1):18, March 2017. doi: 10.3847/1538-4357/aa62a8.

N. Gyenge, T. Singh, T. S. Kiss, A. K. Srivastava, and R. Erdélyi. Active longitude and coronal mass ejection occurrences. *The Astrophysical Journal*, 838(1):18, mar 2017. doi: 10.3847/1538-4357/aa62a8. URL https://dx.doi.org/10.3847/1538-4357/aa62a8.

Susan Haack and William Duica. Evidence and inquiry. towards reconstruction in epistemology. *Ideas y Valores*, 1993.

A Haas, G Matheron, and J Serra. Morphologie mathématique et granulométries en place: Annales mines, v. 11. *1967b, Morphologie mathématique et granulométries en place: Annales Mines*, 12:767–782, 1967.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in Transformer. *arXiv e-prints*, art. arXiv:2103.00112, February 2021. doi: 10.48550/arXiv.2103.00112.

W. Hanke, V. Dose, and U. von Toussaint. Bayesian probability theory, 06 2014. URL https://doi.org/10.1017/cbo9781139565608.

R. A. Harrison, P. Bryans, G. M. Simnett, and M. Lyons. Coronal dimming and the coronal mass ejection onset. *Astronomy and Astrophysics*, 400:1071–1083, March 2003. doi: 10.1051/0004-6361:20030088.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97. URL https://doi.org/10.1093/biomet/57.1.97.

David H. Hathaway. The Solar Cycle. *Living Reviews in Solar Physics*, 7(1):1, December 2010. doi: 10.12942/lrsp-2010-1.

David H. Hathaway and Robert M. Wilson. What the Sunspot Record Tells Us About Space Climate. *Solar Physics*, 224(1-2):5–19, October 2004. doi: 10.1007/s11207-005-3996-8.

Phillip Hess and Jie Zhang. A study of the earth-affecting cmes of solar cycle 24. *Solar Physics*, 292(6):1–20, 2017.

T. Hirayama. Theoretical Model of Flares and Prominences. I: Evaporating Flare Model. *Solar Physics*, 34(2):323–338, February 1974. doi: 10.1007/BF00153671.

Chuanpeng Hou, Jiansen He, Die Duan, Ziqi Wu, Yajie Chen, Daniel Verscharen, Alexis P. Rouillard, Huichao Li, Liping Yang, and Stuart D. Bale. The origin of interplanetary switchbacks in reconnection at chromospheric network boundaries. *Nature Astronomy*, 8(10):1246–1256, October 2024. doi: 10.1038/s41550-024-02321-9.

R. A. Howard, A. Vourlidas, R. C. Colaninno, C. M. Korendyke, S. P. Plunkett, M. T. Carter, D. Wang, N. Rich, S. Lynch, A. Thurn, D. G. Socker, A. F. Thernisien, D. Chua, M. G. Linton, S. Koss, S. Tun-Beltran, H. Dennison, G. Stenborg, D. R. McMullin, T. Hunt, R. Baugh, G. Clifford, D. Keller, J. R. Janesick, J. Tower, M. Grygon, R. Farkas, R. Hagood, K. Eisenhauer, A. Uhl,

S. Yerushalmi, L. Smith, P. C. Liewer, M. C. Velli, J. Linker, V. Bothmer, P. Rochus, J. P. Halain, P. L. Lamy, F. Auchère, R. A. Harrison, A. Rouillard, S. Patsourakos, O. C. St. Cyr, H. Gilbert, H. Maldonado, C. Mariano, and J. Cerullo. The Solar Orbiter Heliospheric Imager (SoloHI). *Astronomy and Astrophysics*, 642:A13, October 2020. doi: 10.1051/0004-6361/201935202.

H. S. Hudson. Solar flares, microflares, nanoflares, and coronal heating. *Solar Physics*, 133(2):357–369, June 1991. doi: 10.1007/BF00149894.

A. Isavnin. FRiED: A Novel Three-dimensional Model of Coronal Mass Ejections. *Astrophysical Journal*, 833(2):267, December 2016. doi: 10.3847/1538-4357/833/2/267.

A. Isavnin, A. Vourlidas, and E. K. J. Kilpua. Three-Dimensional Evolution of Erupted Flux Ropes from the Sun (2 - 20 $R_\odot$) to 1 AU. *Solar Physics*, 284(1):203–215, May 2013. doi: 10.1007/s11207-012-0214-3.

Željko Ivezić, Andrew J Connolly, Jacob T VanderPlas, and Alexander Gray. *Statistics, data mining, and machine learning in astronomy: A practical python guide for the analysis of survey data*. Princeton University Press, 2019.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

Ian T. Jolliffe. Principal component analysis, 04 2005. URL https://doi.org/10.1002/0470013192.bsa501.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments, 04 2016. URL https://doi.org/10.1098/rsta.2015.0202.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, June 2021. doi: 10.1038/s42254-021-00314-5.

J. T. Karpen, S. E. M. Tanner, S. K. Antiochos, and C. R. DeVore. Prominence formation by thermal nonequilibrium in the sheared-arcade model. *The Astrophysical Journal*, 635(2):1319–1328, December 2005. doi: 10.1086/497531. URL https://doi.org/10.1086/497531.

C. Kay, M. Opher, and R. M. Evans. Global Trends of CME Deflections Based on CME and Solar Parameters. *Astrophysical Journal*, 805(2):168, June 2015. doi: 10.1088/0004-637X/805/2/168.

Kartav Kesri, Sahel Dey, Piyali Chatterjee, and Robertus Erdelyi. Dependence of Spicule Properties on the Magnetic Field—Results from Magnetohydrodynamics Simulations. *Astrophysical Journal*, 973(1):49, September 2024. doi: 10.3847/1538-4357/ad67d8.

John Maynard Keynes. *A treatise on probability*. Courier Corporation, 2013.

T. S. Kiss and R. Erdélyi. On Quasi-biennial Oscillations in Chromospheric Macrospicules and Their Potential Relation to the Global Solar Magnetic Field. *Astrophysical Journal*, 857(2):113, April 2018. doi: 10.3847/1538-4357/aab8f7.

T. S. Kiss, N. Gyenge, and R. Erdélyi. Systematic Variations of Macrospicule Properties Observed by SDO/AIA over Half a Decade. *Astrophysical Journal*, 835(1):47, January 2017. doi: 10.3847/1538-4357/aa5272.

Eric W. Koch and Erik W. Rosolowsky. Filament identification through mathematical morphology. *Monthly Notices of the Royal Astronomical Society*, 452(4):3435–3450, 08 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv1521. URL https://doi.org/10.1093/mnras/stv1521.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603638.

Andreĭ Nikolaevich Kolmogorov and Albert T Bharucha-Reid. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.

R. A. Kopp and G. W. Pneuman. Magnetic reconnection in the corona and the loop prominence phenomenon. *Solar Physics*, 50(1):85–98, October 1976. doi: 10.1007/BF00206193.

K. Kravvaris, P. Navrátil, S. Quaglioni, C. Hebborn, and G. Hupin. Ab initio informed evaluation of the radiative capture of protons on ^{7}Be. *Physics Letters B*, 845:138156, October 2023. doi: 10.1016/j.physletb.2023.138156.

K. Kusano, Yumi Bamba, T. Yamamoto, Yusuke Iida, Shin Toriumi, and Ayumi Asai. Magnetic field structures triggering solar flares and coronal mass ejections. *IOP Publishing*, 760(1):31–31, 10 2012. doi: 10.1088/0004-637x/760/1/31. URL https://doi.org/10.1088/0004-637x/760/1/31.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Louis J. Lanzerotti. Space weather effects on technologies. *Washington DC American Geophysical Union Geophysical Monograph Series*, 125:11–22, January 2001. doi: 10.1029/GM125p0011.

Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning, 05 2015. URL https://doi.org/10.1038/nature14539.

James R. Lemen, Alan M. Title, David J. Akin, Paul F. Boerner, Catherine Chou, Jerry F. Drake, Dexter W. Duncan, Christopher G. Edwards, Frank M. Friedlaender, Gary F. Heyman, Neal E. Hurlburt, Noah L. Katz, Gary D. Kushner, Michael Levay, Russell W. Lindgren, Dnyanesh P. Mathur, Edward L. McFeaters, Sarah Mitchell, Roger A. Rehse, Carolus J. Schrijver, Larry A. Springer, Robert A. Stern, Theodore D. Tarbell, Jean-Pierre Wuelser, C. Jacob Wolfson, Carl Yanari, Jay A. Bookbinder, Peter N. Cheimets, David Caldwell, Edward E. Deluca, Richard Gates, Leon Golub, Sang Park, William A. Podgorski, Rock I. Bush, Philip H. Scherrer, Mark A. Gummin, Peter Smith, Gary Auker, Paul Jerram, Peter Pool, Regina Soufli, David L. Windt, Sarah Beardsley, Matthew Clapp, James Lang, and Nicholas Waltham. The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). *Solar Physics*, 275(1-2):17–40, January 2012. doi: 10.1007/s11207-011-9776-8.

Xiaohong Li, Jun Zhang, Shuhong Yang, Yijun Hou, and Robert Erdélyi. Observing Kelvin-Helmholtz instability in solar blowout jet. *Scientific Reports*, 8:8136, May 2018. doi: 10.1038/s41598-018-26581-4.

Yuxi Li. Deep reinforcement learning: An overview, 01 2017. URL https://arxiv.org/abs/1701.07274.

Chia-Hsien Lin and James Chen. Drag force on coronal mass ejections (cmes). *Journal of Geophysical Research: Space Physics*, 127(6):e2020JA028744, 2022. doi: https://doi.org/10.1029/2020JA028744. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JA028744. e2020JA028744 2020JA028744.

R. Lionello, Tibor Török, V. S. Titov, J. E. Leake, Z. Mikić, J. A. Linker, and M. G. Linton. The contribution of coronal jets to the solar wind. *IOP Publishing*, 831(1):L2–L2, 10 2016. doi: 10.3847/2041-8205/831/1/l2. URL https://doi.org/10.3847/2041-8205/831/1/l2.

Jiajia Liu, R. Erdélyi, Yuming Wang, and Rui Liu. Untwisting jets related to magnetic flux cancellation. *IOP Publishing*, 852(1):10–10, 12 2017. doi: 10.3847/1538-4357/aa992d. URL https://doi.org/10.3847/1538-4357/aa992d.

Jiajia Liu, Yudong Ye, Chenglong Shen, Yuming Wang, and Robert Erdélyi. A New Tool for CME Arrival Time Prediction using Machine Learning Algorithms: CAT-PUMA. *Astrophysical Journal*, 855(2):109, March 2018. doi: 10.3847/1538-4357/aaae69.

Jiajia Liu, Anchuan Song, David B. Jess, Jie Zhang, Mihalis Mathioudakis, Szabolcs Soós, Francis P. Keenan, Yuming Wang, and Robertus Erdélyi. Power-law distribution of solar cycle–modulated coronal jets. *The Astrophysical Journal Supplement Series*, 266(1):17, may 2023. doi: 10.3847/1538-4365/acc85a. URL https://dx.doi.org/10.3847/1538-4365/acc85a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

R. M. MacQueen and R. R. Fisher. The Kinematics of Solar Inner Coronal Transients. *Solar Physics*, 89(1):89–102, November 1983. doi: 10.1007/BF00211955.

Anwesha Maharana, Alexey Isavnin, Camilla Scolini, Nicolas Wijsen, Luciano Rodriguez, Marilena Mierla, Jasmina Magdalenić, and Stefaan Poedts. Implementation and validation of the FRi3D flux rope model in EUHFORIA. *Advances in Space Research*, 70(6):1641–1662, September 2022. doi: 10.1016/j.asr.2022.05.056.

Pierre Simon Marquis de Laplace. *A philosophical essay on probabilities*. Wiley, 1902.

Georges Matheron. *Eléments pour une théorie des milieux poreux*. Masson, Paris, 1967.

R. L. McPherron and J. M. Weygand. The solar wind and geomagnetic activity as a function of time relative to corotating interaction regions, 01 2006. URL https://doi.org/10.1029/167gm12.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Z. Mikić and M. A. Lee. An Introduction to Theory and Models of CMEs, Shocks, and Solar Energetic Particles. *Space Science Reviews*, 123(1-3):57–80, March 2006. doi: 10.1007/s11214-006-9012-2.

Richard v Mises. *Wahrscheinlichkeit Statistik und Wahrheit*, volume 7. Springer-Verlag, 2013.

Rory Mitchell and Eibe Frank. Accelerating the xgboost algorithm using gpu computing. *PeerJ, Inc.*, 3:e127–e127, 07 2017. doi: 10.7717/peerj-cs.127. URL https://doi.org/10.7717/peerj-cs.127.

Neelam Mittal and U. Narain. Initiation of cmes: A review, 06 2010. URL https://doi.org/10.1016/j.jastp.2010.03.011.

Ranjan Mondal, Moni Shankar Dey, and Bhabatosh Chanda. Image restoration by learning morphological opening-closing network. *Mathematical Morphology - Theory and Applications*, 4(1):87–107, 2020. doi: doi:10.1515/mathm-2020-0103. URL https://doi.org/10.1515/mathm-2020-0103.

Ronald L. Moore, Jonathan W. Cirtain, Alphonse C. Sterling, and David A. Falconer. Dichotomy of Solar Coronal Jets: Standard Jets and Blowout Jets. *Astrophysical Journal*, 720(1):757–770, September 2010a. doi: 10.1088/0004-637X/720/1/757.

Ronald L. Moore, Jonathan W. Cirtain, Alphonse C. Sterling, and David A. Falconer. Dichotomy of Solar Coronal Jets: Standard Jets and Blowout Jets. *Astrophysical Journal*, 720(1):757–770, September 2010b. doi: 10.1088/0004-637X/720/1/757.

R. J. Morton, A. K. Srivastava, and R. Erdélyi. Observations of quasi-periodic phenomena associated with a large blowout solar jet. *Astronomy and Astrophysics*, 542:A70, June 2012. doi: 10.1051/0004-6361/201117218.

Ronish Mugatwala, Simone Chierichini, Gregoire Francisco, Gianluca Napoletano, Raffaello Foldes, Luca Giovannelli, Giancarlo De Gasperis, Enrico Camporeale, Robertus Erdélyi, and Dario Del Moro. A catalogue of observed geo-effective CME/ICME characteristics. *Journal of Space Weather and Space Climate*, 14:6, February 2024. doi: 10.1051/swsc/2024004.

N. Muhr, A. M. Veronig, I. W. Kienreich, M. Temmer, and B. Vršnak. Analysis of Characteristic Parameters of Large-scale Coronal Waves Observed by the Solar-Terrestrial Relations Observatory/Extreme Ultraviolet Imager. *Astrophysical Journal*, 739(2):89, October 2011. doi: 10.1088/0004-637X/739/2/89.

K. Murawski, Z. E. Musielak, and D. Wójcik. 3D Numerical Simulations of Solar Quiet Chromosphere Wave Heating. *The Astrophysical Journal Letters*, 896(1):L1, June 2020. doi: 10.3847/2041-8213/ab94a9.

Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley*, 2(1):86–97, 12 2011. doi: 10.1002/widm.53. URL https://doi.org/10.1002/widm.53.

Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. An introduction to decision tree modeling. *Wiley*, 18(6):275–285, 06 2004. doi: 10.1002/cem.873. URL https://doi.org/10.1002/cem.873.

Gianluca Napoletano, Roberta Forte, Dario Del Moro, Ermanno Pietropaolo, Luca Giovannelli, and Francesco Berrilli. A probabilistic approach to the drag-based model. *J. Space Weather Space Clim.*, 8:A11, 2018a. doi: 10.1051/swsc/2018003. URL https://doi.org/10.1051/swsc/2018003.

Gianluca Napoletano, Roberta Forte, Dario Del Moro, Ermanno Pietropaolo, Luca Giovannelli, and Francesco Berrilli. A probabilistic approach to the drag-based model. *J. Space Weather Space Clim.*, 8:A11, 2018b. doi: 10.1051/swsc/2018003. URL https://doi.org/10.1051/swsc/2018003.

Gianluca Napoletano, Raffaello Foldes, Enrico Camporeale, Giancarlo de Gasperis, Luca Giovannelli, Evangelos Paouris, Ermanno Pietropaolo,

Jannis Teunissen, Ajay Kumar Tiwari, and Dario Del Moro. Parameter distributions for the drag-based modeling of cme propagation. *Space Weather*, page e2021SW002925, 2022. doi: 10.1029/2021SW002925.

Issam El Naqa and Martin J. Murphy. What is machine learning?, 01 2015. URL https://doi.org/10.1007/978-3-319-18305-3_1.

Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers Media*, 7, 01 2013. doi: 10.3389/fnbot.2013.00021. URL https://doi.org/10.3389/fnbot.2013.00021.

Arundhati Navada, Aamir Nizam Ansari, Siddharth Patil, and Balwant A. Sonkamble. Overview of use of decision tree algorithms in machine learning. In *2011 IEEE Control and System Graduate Research Colloquium*, pages 37–42, 2011. doi: 10.1109/ICSGRC.2011.5991826.

Giuseppe Nisticò, V. Bothmer, S. Patsourakos, and G. Zimbardo. Characteristics of euv coronal jets observed with stereo/secchi. *Springer Science+Business Media*, 259(1-2):87–108, 09 2009. doi: 10.1007/s11207-009-9424-8. URL https://doi.org/10.1007/s11207-009-9424-8.

Keiller Nogueira, Jocelyn Chanussot, Mauro Dalla Mura, and Jefersson A. dos Santos. An introduction to deep morphological networks, 2019. URL https://arxiv.org/abs/1906.01751.

Åke Nordlund, Robert F Stein, and Martin Asplund. Solar surface convection. *Living Reviews in Solar Physics*, 6(1):1–117, 2009. doi: 10.12942/lrsp-2009-2.

D Odstrcil. Modeling 3-d solar wind structure. *Advances in Space Research*, 32 (4):497–506, 2003.

Y. Ogawara. Yohkoh (Solar-A) observations of solar activity. *Journal of Atmospheric and Terrestrial Physics*, 57(12):1361–1368, October 1995. doi: 10.1016/0021-9169(94)00137-D.

T. Ohmi, M. Kojima, M. Tokumaru, K. Fujiki, and K. Hakamada. Origin of the slow solar wind. *Advances in Space Research*, 33(5):689–695, January 2004. doi: 10.1016/S0273-1177(03)00238-2.

M. J. Owens and M. Grandé. Predictions of the arrival time of coronal mass ejections at 1au: an analysis of the causes of errors. *Copernicus Publications*, 22(2):661–671, 01 2004. doi: 10.5194/angeo-22-661-2004. URL https://doi.org/10.5194/angeo-22-661-2004.

Mathew Owens and P Cargill. Predictions of the arrival time of coronal mass ejections at 1au: an analysis of the causes of errors. In *Annales Geophysicae*, volume 22, pages 661–671. Copernicus GmbH, 2004.

Mathew James Owens, PJ Cargill, C Pagel, GL Siscoe, and NU Crooker. Characteristic magnetic field and speed properties of interplanetary coronal mass ejections and their sheath regions. *Journal of Geophysical Research: Space Physics*, 110(A1), 2005.

Erika Palmerio, Christina O. Lee, Ian G. Richardson, Teresa Nieves-Chinchilla, Luiz F. G. Dos Santos, Jacob R. Gruesbeck, Nariaki V. Nitta, M. Leila Mays, Jasper S. Halekas, Cary Zeitlin, Shaosui Xu, Mats Holmström, Yoshifumi Futaana, Tamitha Mulligan, Benjamin J. Lynch, and Janet G. Luhmann. CME Evolution in the Structured Heliosphere and Effects at Earth and Mars During Solar Minimum. *Space Weather*, 20(9):e2022SW003215, September 2022. doi: 10.1029/2022SW003215.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

V. Pant, S. Majumdar, R. Patel, A. Chauhan, D. Banerjee, and N. Gopalswamy. Investigating width distribution of slow and fast CMEs in solar cycles 23 and 24. *Frontiers in Astronomy and Space Sciences*, 8:73, May 2021. doi: 10.3389/fspas.2021.634358.

Evangelos Paouris, Jaša Čalogović, Mateja Dumbović, M Leila Mays, Angelos Vourlidas, Athanasios Papaioannou, Anastasios Anastasiadis, and Georgios Balasis. Propagating conditions and the time of icme arrival: A comparison of the effective acceleration model with enlil and dbem models. *Solar Physics*, 296(1):12, 2021. doi: 10.1007/s11207-020-01747-4.

E. Pariat, K. Dalmasse, C. R. DeVore, S. K. Antiochos, and J. T. Karpen. Model for straight and helical solar jets. I. Parametric studies of the magnetic field geometry. *Astronomy and Astrophysics*, 573:A130, January 2015. doi: 10.1051/0004-6361/201424209.

E. Pariat, K. Dalmasse, C. R. DeVore, S. K. Antiochos, and J. T. Karpen. A model for straight and helical solar jets. II. Parametric study of the plasma beta. *Astronomy and Astrophysics*, 596:A36, November 2016. doi: 10.1051/0004-6361/201629109.

E. N. Parker. Dynamics of the Interplanetary Gas and Magnetic Fields. *Astrophysical Journal*, 128:664, November 1958. doi: 10.1086/146579.

S Patsourakos, A Vourlidas, T Török, B Kliem, SK Antiochos, V Archontis, G Aulanier, X Cheng, G Chintzoglou, MK Georgoulis, et al. Decoding the pre-eruptive magnetic field configurations of coronal mass ejections. *Space Science Reviews*, 216(8):1–63, 2020.

Mirko Piersanti, Paola De Michelis, Dario Del Moro, Roberta Tozzi, Michael Pezzopane, Giuseppe Consolini, Maria Federica Marcucci, Monica Laurenza, Simone Di Matteo, Alessio Pignalberi, Virgilio Quattrociocchi, and Piero Diego. From the Sun to Earth: effects of the 25 August 2018 geomagnetic storm. *Annales Geophysicae*, 38(3):703–724, June 2020. doi: 10.5194/angeo-38-703-2020.

L. A. Plyusnina. Determination of the Rotation Periods of Solar Active Longitudes. *Solar Physics*, 261(2):223–232, February 2010. doi: 10.1007/s11207-009-9501-z.

Jens Pomoell and Stefaan Poedts. Euhforia: European heliospheric forecasting information asset. *Journal of Space Weather and Space Climate*, 8:A35, 2018.

E. R. Priest and T. G. Forbes. The magnetic nature of solar flares. *Springer Science+Business Media*, 10(4):313–377, 07 2001. doi: 10.1007/s001590100013. URL https://doi.org/10.1007/s001590100013.

Stefano Pucci, Giannina Poletto, Alphonse C. Sterling, and Marco Romoli. Solar Polar X-Ray Jets and Multiple Bright Points: Evidence for Sympathetic Activity. *The Astrophysical Journal Letters*, 745(2):L31, February 2012. doi: 10.1088/2041-8205/745/2/L31.

Tuija Pulkkinen. Space weather: terrestrial perspective. *Living Reviews in Solar Physics*, 4(1):1–60, 2007.

F. P. Ramunno, S. Hackstein, V. Kinakh, M. Drozdova, G. Quétant, A. Csillaghy, and S. Voloshynovskiy. Solar synthetic imaging: Introducing denoising diffusion probabilistic models on SDO/AIA data. *Astronomy and Astrophysics*, 686:A285, June 2024. doi: 10.1051/0004-6361/202347860.

P. V. S. Rama Rao, S. Gopi Krishna, J.V.R. Prasad, S. Prasad, D. S. V. V. D. Prasad, and K. Niranjan. Geomagnetic storm effects on gps based navigation. *Copernicus Publications*, 27(5):2101–2110, 05 2009. doi: 10.5194/angeo-27-2101-2009. URL https://doi.org/10.5194/angeo-27-2101-2009.

N. E. Raouafi, S. Patsourakos, E. Pariat, P. R. Young, A. C. Sterling, A. Savcheva, M. Shimojo, F. Moreno-Insertis, C. R. DeVore, V. Archontis, T. Török, H. Mason, W. Curdt, K. Meyer, K. Dalmasse, and Y. Matsui. Solar Coronal Jets: Observations, Theory, and Modeling. *Space Science Reviews*, 201(1-4): 1–53, November 2016. doi: 10.1007/s11214-016-0260-5.

N. E. Raouafi, S. Patsourakos, E. Pariat, Peter R. Young, Alphonse C. Sterling, Antonia Savcheva, M. Shimojo, F. Moreno-Insertis, C. R. DeVore, V. Archontis, Tibor Török, H. E. Mason, W. Curdt, Karen Meyer, K. Dalmasse, and Yutaka Matsui. Solar coronal jets: Observations, theory, and modeling. *Springer Science+Business Media*, 201(1-4):1–53, 07 2016. doi: 10.1007/s11214-016-0260-5. URL https://doi.org/10.1007/s11214-016-0260-5.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_565. URL https://doi.org/10.1007/978-0-387-39940-9_565.

Marco Regis. Introduzione alla statistica bayesiana, 2015.

Markus Reichstein, Zavud Baghirov, Martin Jung, and Basil Kraft. Deep learning and Process Understanding for Data-Driven Earth System Science. In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, page 15874, April 2024. doi: 10.5194/egusphere-egu24-15874.

Ian G Richardson and Hilary V Cane. Near-earth interplanetary coronal mass ejections during solar cycle 23 (1996–2009): Catalog and summary of properties. *Solar Physics*, 264(1):189–237, 2010.

Christian P Robert et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.

Lior Rokach. Decision forest: Twenty years of research. *Elsevier BV*, 27:111–125, 01 2016. doi: 10.1016/j.inffus.2015.06.005. URL https://doi.org/10.1016/j.inffus.2015.06.005.

A. P. Rouillard, J. A. Davies, R. J. Forsyth, A. Rees, C. J. Davis, R. A. Harrison, M. Lockwood, D. Bewsher, S. R. Crothers, C. J. Eyles, Mike Hapgood, and C. H. Perry. First imaging of corotating interaction regions using the stereo spacecraft. *American Geophysical Union*, 35(10), 05 2008. doi: 10.1029/2008gl033767. URL https://doi.org/10.1029/2008gl033767.

A. P. Rouillard, I. Plotnikov, R. F. Pinto, M. Tirole, M. Lavarra, P. Zucca, R. Vainio, A. J. Tylka, A. Vourlidas, M. L. De Rosa, J. Linker, A. Warmuth, G. Mann, C. M. S. Cohen, and R. A. Mewaldt. Deriving the Properties of Coronal Pressure Fronts in 3D: Application to the 2012 May 17 Ground Level Enhancement. *Astrophysical Journal*, 833(1):45, December 2016. doi: 10.3847/1538-4357/833/1/45.

Mrinmoy Roy, Sarwar J. Minar, Porarthi Dhar, and Ahmad Faruq. Machine learning applications in healthcare: The state of knowledge and future directions, 01 2023. URL https://arxiv.org/abs/2307.14067.

Swalpa Roy, Ranjan Mondal, Mercedes Paoletti, Juan Haut, and Antonio Plaza. Morphological convolutional neural networks for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:8689–8702, 06 2021. doi: 10.1109/JSTARS.2021.3088228.

A. Ruffenach, B. Lavraud, C. J. Farrugia, P. Démoulin, S. Dasso, M. J. Owens, J. A. Sauvaud, A. P. Rouillard, A. Lynnyk, C. Foullon, N. P. Savani, J. G. Luhmann, and A. B. Galvin. Statistical study of magnetic cloud erosion by magnetic reconnection. *Journal of Geophysical Research (Space Physics)*, 120(1): 43–60, January 2015. doi: 10.1002/2014JA020628.

Ranadeep Sarkar, Nandita Srivastava, Marilena Mierla, Matthew J. West, and Elke D'Huys. Evolution of the Coronal Cavity From the Quiescent to Eruptive Phase Associated with Coronal Mass Ejection. *Astrophysical Journal*, 875(2):101, April 2019. doi: 10.3847/1538-4357/ab11c5.

Iqbal H. Sarker. Machine learning: Algorithms, real-world applications and research directions, 03 2021. URL https://doi.org/10.1007/s42979-021-00592-x.

A. Savcheva, E. Pariat, S. McKillop, P. McCauley, E. Hanson, Y. Su, E. Werner, and E. E. DeLuca. The Relation between Solar Eruption Topologies and

Observed Flare Features. I. Flare Ribbons. *Astrophysical Journal*, 810(2):96, September 2015. doi: 10.1088/0004-637X/810/2/96.

Brigitte Schmieder, Reetika Joshi, and Ramesh Chandra. Solar jets observed with the Interface Region Imaging Spectrograph (IRIS). *Advances in Space Research*, 70(6):1580–1591, September 2022. doi: 10.1016/j.asr.2021.12.013.

Bernhard Schölkopf, Christopher JC Burges, and Alexander J Smola. *Advances in kernel methods: support vector learning*. MIT press, 1999. doi: 10.7551/mitpress/1130.003.0001.

Carolus J. Schrijver and George L. Siscoe. *Heliophysics: Space Storms and Radiation: Causes and Effects*. Cambridge University Press, 2010.

Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited. *Association for Computing Machinery*, 42(3):1–21, 07 2017. doi: 10.1145/3068335. URL https://doi.org/10.1145/3068335.

Rainer Schwenn. Space weather: The solar perspective. *Living reviews in solar physics*, 3(1):1–72, 2006.

Camilla Scolini and Erika Palmerio. The spheroid CME model in EUHFORIA. *Journal of Space Weather and Space Climate*, 14:13, May 2024. doi: 10.1051/swsc/2024011.

E. Scullion, L. Rouppe van der Voort, and J. de la Cruz Rodriguez. Type-II spicules: Heating and magnetic field properties from aligned CRISP/SST and SDO observations. In *SDO-4: Dynamics and Energetics of the Coupled Solar Atmosphere. The Synergy Between State-of-the-Art Observations and Numerical Simulations*, page 44, March 2012.

Terrence J. Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. *National Academy of Sciences*, 117(48):30033–30038, 01 2020. doi: 10.1073/pnas.1907373117. URL https://doi.org/10.1073/pnas.1907373117.

Jaydip Sen, Rajdeep Sen, and Abhishek Dutta. Machine learning in finance-emerging trends and challenges, 01 2021. URL https://arxiv.org/abs/2110.11999.

J. Serra. *Introduction à la morphologie mathématique*. Cahiers du Centre de morphologie mathématique de Fontainebleau. Centre de morphologie mathématique de Fontainebleau, Fontainebleau, 1969. URL https://books.google.fr/books?id=5dNcPgAACAAJ.

J Serra. Image analysis and mathematical morphology, 1983.

NR Sheeley Jr, JH Walters, Y-M Wang, and RA Howard. Continuous tracking of coronal outflows: Two kinds of coronal mass ejections. *Journal of Geophysical Research: Space Physics*, 104(A11):24739–24767, 1999.

Chenglong Shen, Yuming Wang, Zonghao Pan, Min Zhang, Pinzhong Ye, and S Wang. Full halo coronal mass ejections: Do we need to correct the projection effect in terms of velocity? *Journal of Geophysical Research: Space Physics*, 118(11):6858–6865, 2013.

Bidzina M. Shergelashvili, Stefaan Poedts, and Avtandil D. Pataraya. Non-modal Cascade in the Compressible Solar Atmosphere: Self-Heating, an Alternative Way to Enhance Wave Heating. *The Astrophysical Journal Letters*, 642(1):L73–L76, May 2006. doi: 10.1086/504350.

Tong Shi, Yikang Wang, Linfeng Wan, Xin Cheng, Mingde Ding, and Jie Zhang. Predicting the arrival time of coronal mass ejections with the graduated cylindrical shell and drag force model. *The Astrophysical Journal*, 806(2):271, 2015. doi: 10.1088/0004-637X/806/2/271.

Kazunari Shibata, Yoshinori Ishido, Loren W. Acton, Keith T. Strong, Tadashi Hirayama, Yutaka Uchida, Alan H. McAllister, Ryoji Matsumoto, Saku Tsuneta, Toshifumi Shimizu, Hirohisa Hara, Takashi Sakurai, Kiyoshi Ichimoto, Yohei Nishino, and Yoshiaki Ogawara. Observations of X-Ray Jets with the YOHKOH Soft X-Ray Telescope. *Publications of the Astronomical Society of Japan*, 44:L173–L179, October 1992.

Kazunari Shibata, Tahei Nakamura, Takuma Matsumoto, Kenichi Otsuji, Takenori J. Okamoto, Naoto Nishizuka, Tomoko Kawate, Hiroko Watanabe, Shin'ichi Nagata, Satoru UeNo, Reizaburo Kitai, Satoshi Nozawa, Saku Tsuneta, Yoshinori Suematsu, Kiyoshi Ichimoto, Toshifumi Shimizu, Yukio Katsukawa, Theodore D. Tarbell, Thomas E. Berger, Bruce W. Lites, Richard A. Shine, and Alan M. Title. Chromospheric Anemone Jets as Evidence of Ubiquitous Reconnection. *Science*, 318(5856):1591, December 2007. doi: 10.1126/science.1146708.

Frank Y. Shih and Artur J. Kowalski. Automatic Extraction of Filaments in Hα Solar Images. *Solar Physics*, 218(1):99–122, December 2003. doi: 10.1023/B:SOLA.0000013052.34180.58.

Masumi Shimojo, Shizuyo Hashimoto, Kazunari Shibata, Tadashi Hirayama, Hugh S. Hudson, and Loren W. Acton. Statistical Study of Solar X-Ray Jets Observed with the YOHKOH Soft X-Ray Telescope. *Publications of the Astronomical Society of Japan*, 48:123–136, February 1996. doi: 10.1093/pasj/48.1.123.

Ashok K. Singh, Devendraa Siingh, and R. P. Singh. Space weather: Physics, effects and predictability. *Springer Science+Business Media*, 31(6):581–638, 10 2010. doi: 10.1007/s10712-010-9103-1. URL https://doi.org/10.1007/s10712-010-9103-1.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14:199–222, 2004. doi: 10.1023/B:STCO.0000035301.49549.88.

P Soille. Morphological image analysis: Principles and applications, 1999.

Sami K. Solanki and Natalie A. Krivova. Analyzing Solar Cycles. *Science*, 334 (6058):916, November 2011. doi: 10.1126/science.1212555.

Sami K. Solanki, Bernd Inhester, and Manfred Schüssler. The solar magnetic field. *Reports on Progress in Physics*, 69(3):563–668, March 2006. doi: 10.1088/0034-4885/69/3/R02.

Charles P Sonett. A summary review of the scientific findings of the mariner venus mission. *Space Science Reviews*, 2(6):751–777, 1963.

H. Q. Song, Z. Zhong, Y. Chen, J. Zhang, X. Cheng, L. Zhao, Q. Hu, and G. Li. A Statistical Study of the Average Iron Charge State Distributions inside Magnetic Clouds for Solar Cycle 23. *The Astrophysical Journal Supplement Series*, 224(2):27, June 2016. doi: 10.3847/0067-0049/224/2/27.

Sz. Soós, J. Liu, M. B. Korsós, and R. Erdélyi. Evolution of Coronal Jets during Solar Cycle 24. *Astrophysical Journal*, 965(1):43, April 2024. doi: 10.3847/1538-4357/ad29f8.

K. G. Srinivasa, Siddesh Gaddadevara Matt, and H. Srinidhi. Basics of machine learning, 01 2018. URL https://doi.org/10.1007/978-3-319-77800-6_8.

Sotiris Stamkos, Spiros Patsourakos, Angelos Vourlidas, and Ioannis A. Daglis. How Magnetic Erosion Affects the Drag-Based Kinematics of Fast Coronal Mass Ejections. *Solar Physics*, 298(7):88, July 2023. doi: 10.1007/s11207-023-02178-7.

Michael Steinbach, Levent Ertöz, and Vipin Kumar. The challenges of clustering high dimensional data, 01 2004. URL https://doi.org/10.1007/978-3-662-08968-2_16.

Alphonse C. Sterling and Ronald L. Moore. Coronal-jet-producing Minifilament Eruptions as a Possible Source of Parker Solar Probe Switchbacks. *The Astrophysical Journal Letters*, 896(2):L18, June 2020. doi: 10.3847/2041-8213/ab96be.

Alphonse C. Sterling, Ronald L. Moore, David A. Falconer, and Mitzi Adams. Small-scale filament eruptions as the driver of X-ray jets in solar coronal holes. *Nature*, 523(7561):437–440, July 2015. doi: 10.1038/nature14556.

Michael Stix. *The Sun. an Introduction*. Springer Science & Business Media, 1989.

P. A. Sturrock. Model of the High-Energy Phase of Solar Flares. *Nature*, 211 (5050):695–697, August 1966. doi: 10.1038/211695a0.

Prasad Subramanian, Alejandro Lara, and Andrea Borgazzi. Can solar wind viscous drag account for coronal mass ejection deceleration? *Geophysical Research Letters*, 39(19), 2012. doi: https://doi.org/10.1029/2012GL053625. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012GL053625.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A Survey on Deep Transfer Learning. *arXiv e-prints*, art. arXiv:1808.01974, August 2018. doi: 10.48550/arXiv.1808.01974.

Manuela Temmer. Space weather: the solar perspective–an update to schwenn (2006). *arXiv preprint arXiv:2104.04261*, 2021.

Adithya Thaduri, Diego Galar, and Uday Kumar. Space weather climate impacts on railway infrastructure. *International Journal of System Assurance Engineering and Management*, 11:267–281, 2020. doi: 10.1007/s13198-020-01003-9.

B. J. Thompson, S. P. Plunkett, J. B. Gurman, J. S. Newmark, O. C. St. Cyr, and D. J. Michels. SOHO/EIT observations of an Earth-directed coronal mass ejection on May 12, 1997. *Geophysical Research Letters*, 25(14):2465–2468, July 1998. doi: 10.1029/98GL50429.

Neil Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. The computational limits of deep learning, 01 2020. URL https://arxiv.org/abs/2007.05558.

H. Tian, E. E. DeLuca, S. R. Cranmer, B. De Pontieu, H. Peter, J. Martínez-Sykora, L. Golub, S. McKillop, K. K. Reeves, M. P. Miralles, P. McCauley, S. Saar, P. Testa, M. Weber, N. Murphy, J. Lemen, A. Title, P. Boerner, N. Hurlburt, T. D. Tarbell, J. P. Wuelser, L. Kleint, C. Kankelborg, S. Jaeggli, M. Carlsson, V. Hansteen, and S. W. McIntosh. Prevalence of small-scale jets from the networks of the solar transition region and chromosphere. *Science*, 346(6207):1255711, October 2014. doi: 10.1126/science.1255711.

Tibor Török, R. Lionello, V. S. Titov, J. E. Leake, Z. Mikić, J. A. Linker, and M. G. Linton. Modeling jets in the corona and solar wind, 01 2015. URL https://arxiv.org/abs/1511.09350.

Mathew J Owens undefined. Solar-wind structure, 02 2020. URL https://oxfordre.com/physics/display/10.1093/acrefore/9780190871994.001.0001/acrefore-9780190871994-e-19;jsessionid=95860990F4CA8E083B83F28B7117836A.

Vishal Upendran, Mark C. M. Cheung, Shravan Hanasoge, and Ganapathy Krishnamurthi. Solar Wind Prediction Using Deep Learning. *Space Weather*, 18(9):e02478, September 2020. doi: 10.1029/2020SW002478.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *The MIT Press*, 9(86):2579–2605, 01 2008. URL http://isplab.tudelft.nl/sites/default/files/vandermaaten08a.pdf.

Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Springer Science+Business Media*, 109(2):373–440, 11 2019. doi: 10.1007/s10994-019-05855-6. URL https://doi.org/10.1007/s10994-019-05855-6.

Don van Ravenzwaaij, Pete Cassey, and Scott Brown. A simple introduction to markov chain monte–carlo sampling. *Springer Science+Business Media*, 25 (1):143–154, 03 2016. doi: 10.3758/s13423-016-1015-8. URL https://doi.org/10.3758/s13423-016-1015-8.

Luc Vincent and Pierre Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(06):583–598, 1991.

A Vourlidas, S Patsourakos, and NP Savani. Predicting the geoeffective properties of coronal mass ejections: current status, open issues and path forward. *Philosophical Transactions of the Royal Society A*, 377(2148):20180096, 2019.

Bojan Vršnak, T Žic, D Vrbanec, M Temmer, T Rollett, C Möstl, A Veronig, J Čalogović, M Dumbović, S Lulić, et al. Propagation of interplanetary coronal mass ejections: The drag-based model. *Solar physics*, 285(1):295–315, 2013.

Bojan Vršnak, M Temmer, T Žic, A Taktakishvili, M Dumbović, C Möstl, AM Veronig, ML Mays, and D Odstrčil. Heliospheric propagation of coronal mass ejections: comparison of numerical wsa-enlil+ cone model and analytical drag-based model. *The Astrophysical Journal Supplement Series*, 213(2):21, 2014.

B. Vršnak, H. Aurass, J. Magdalenić, and N. Gopalswamy. Band-splitting of coronal and interplanetary type II bursts. I. Basic properties. *Astronomy and Astrophysics*, 377:321–329, October 2001. doi: 10.1051/0004-6361:20011067.

B. Vršnak. Processes and mechanisms governing the initiation and propagation of cmes. *Copernicus Publications*, 26(10):3089–3101, 10 2008. doi: 10.5194/angeo-26-3089-2008. URL https://doi.org/10.5194/angeo-26-3089-2008.

Jing Wang, Kai Yang, and P. F. Chen. Is flux rope a necessary condition for the progenitor of coronal mass ejections? *IOP Publishing*, 815(1):72–72, 12 2015. doi: 10.1088/0004-637x/815/1/72. URL https://doi.org/10.1088/0004-637x/815/1/72.

Yimin Wang, Jiajia Liu, Ye Jiang, and Robert Erdélyi. CME Arrival Time Prediction Using Convolutional Neural Network. *Astrophysical Journal*, 881 (1):15, August 2019. doi: 10.3847/1538-4357/ab2b3e.

Yimin Wang, Jiajia Liu, Ye Jiang, and Robert Erdélyi. Cme arrival time prediction using convolutional neural network. *The Astrophysical Journal*, 881(1):15, 2019.

Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 03 1963. doi: 10.1080/01621459.1963.10500845. URL https://doi.org/10.1080/01621459.1963.10500845.

David F Webb and Timothy A Howard. Coronal mass ejections: Observations. *Living Reviews in Solar Physics*, 9(1):1–83, 2012.

George L. Withbroe. Solar activity cycle - History and predictions. *Journal of Spacecraft and Rockets*, 26:394–402, December 1989. doi: 10.2514/3.26085.

Sanjay Yadav and Sanyam Shukla. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 78–83, 2016. doi: 10.1109/IACC.2016.25.

S Yashiro, N Gopalswamy, G Michalek, OC St. Cyr, SP Plunkett, NB Rich, and RA Howard. A catalog of white light coronal mass ejections observed by the soho spacecraft. *Journal of Geophysical Research: Space Physics*, 109(A7), 2004. doi: 10.1029/2003JA010282.

Peter R. Young and K. Muglach. A coronal hole jet observed with hinode and the solar dynamics observatory. *Oxford University Press*, 66(SP1), 11 2014. doi: 10.1093/pasj/psu088. URL https://doi.org/10.1093/pasj/psu088.

M. Youssef. On the relation between the cmes and the solar flares. *NRIAG Journal of Astronomy and Geophysics*, 1(2):172–178, 2012. ISSN 2090-9977. doi: https://doi.org/10.1016/j.nrjag.2012.12.014. URL https://www.sciencedirect.com/science/article/pii/S2090997712000235.

J Zhang and KP Dere. A statistical study of main and residual accelerations of coronal mass ejections. *The Astrophysical Journal*, 649(2):1100, 2006.

J. Zhang, I. G. Richardson, D. F. Webb, N. Gopalswamy, E. Huttunen, J. C. Kasper, N. V. Nitta, W. Poomvises, B. J. Thompson, C. C. Wu, S. Yashiro, and A. N. Zhukov. Solar and interplanetary sources of major geomagnetic storms (Dst <= -100 nT) during 1996-2005. *Journal of Geophysical Research (Space Physics)*, 112(A10):A10102, October 2007. doi: 10.1029/2007JA012321.

L. Y. Zhang, H. N. Wang, and Z. L. Du. Prediction of solar active longitudes. *A&A*, 484(2):523–527, 2008. doi: 10.1051/0004-6361:200809464. URL https://doi.org/10.1051/0004-6361:200809464.

T. Zhang, Q. Hao, and P. F. Chen. Statistical analyses of solar prominences and active region features in 304 å filtergrams detected via deep learning. *The Astrophysical Journal Supplement Series*, 272, 2024. URL https://api.semanticscholar.org/CorpusID:267770259.

Liang Zhao, Henry Han, Susan T. Lepri, and Ryan Dewey. Classification of in-situ solar wind data measured by solar orbiter/swa-pas and his using machine learning. In Henry Han, editor, *Recent Advances in Next-Generation Data Science*, pages 183–198, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-67871-4.

Ruisheng Zheng, Yao Chen, Bing Wang, Hongqiang Song, and Wenda Cao. Formation of a tiny flux rope in the center of an active region driven by

magnetic flux emergence, convergence, and cancellation. *EDP Sciences*, 642:A199–A199, 10 2020. doi: 10.1051/0004-6361/202037475. URL https://doi.org/10.1051/0004-6361/202037475.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43–76, 2019. URL https://api.semanticscholar.org/CorpusID:207847753.

Rens Zwaard, Matthias Bergmann, Joe Zender, R. Kariyappa, Gabriel Giono, and Luc Damé. Segmentation of coronal features to understand the solar euv and uv irradiance variability iii. inclusion and analysis of bright points. *Solar Physics*, 296, 09 2021. doi: 10.1007/s11207-021-01863-9.

This publication is part of the Space Weather Awareness Training Network (SWATNet) which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Innovative Training Networks, Grant Agreement No 955620. The publication reflects only the author's view and does not represent the opinion of the European Commission (EC), and the EC is not responsible for any use that might be made of information contained.