



OPEN ACCESS

EDITED BY

Weichao Tu,
West Virginia University, United States

REVIEWED BY

Andrew Smith,
Northumbria University, United Kingdom
Yue Chen,
Los Alamos National Laboratory (DOE),
United States

*CORRESPONDENCE

Xuedong Feng,
✉ fengxd2021@mail.sustech.edu.cn

RECEIVED 24 February 2025

ACCEPTED 12 May 2025

PUBLISHED 03 June 2025

CITATION

Feng X, Yang J, Bortnik J, Wang C-P and Liu J
(2025) Predicting characteristics of bursty
bulk flows in Earth's plasma sheet using
machine learning techniques.
Front. Astron. Space Sci. 12:1582607.
doi: 10.3389/fspas.2025.1582607

COPYRIGHT

© 2025 Feng, Yang, Bortnik, Wang and Liu.
This is an open-access article distributed
under the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited,
in accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Predicting characteristics of bursty bulk flows in Earth's plasma sheet using machine learning techniques

Xuedong Feng^{1*}, Jian Yang¹, Jacob Bortnik², Chih-Ping Wang²
and Jiang Liu^{2,3}

¹Department of Earth and Space Sciences, Southern University of Science and Technology, Shenzhen, China, ²Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, Los Angeles, CA, United States, ³Department of Earth, Planetary, and Space Sciences, University of California, Los Angeles, Los Angeles, CA, United States

Bursty bulk flows (BBFs) play a crucial role in transporting energy, mass, and magnetic flux from the Earth's magnetotail to the near-Earth region. However, their impulsive nature and small spatial scale pose significant difficulties for *in-situ* observations, given that only a handful number of spacecraft operate within the vast expanse of the magnetotail. Consequently, accurately predicting their behavior remains a challenging goal. In this study, we employ the XGBoost machine learning algorithm to predict the variation range of several essential BBF properties, including duration, magnetic field, plasma moments, and specific entropy parameters. The observed characteristics of a BBF are shaped by its formation in the downstream tail and its journey until it reaches the spacecraft. Therefore, we use both the background properties of the plasma sheet prior to the arrival of the BBF and the attributes of indirectly related variables during the BBF interval as inputs. Trained on 17 years of THEMIS data, we explore different input configurations. One approach involves incorporating optimal parameter combinations, utilizing as many input parameters as possible to predict upper and lower bounds of a target variable. Within this framework, we further apply the leave-one-feature-out method to quantitatively assess the contribution of each input, identifying the most dominant factor influencing BBFs in a statistical sense. Another approach involves cross-instrument prediction, leveraging measurements from a different payload. Our findings reveal that including observed background values enhances prediction accuracy by 10–20 percentage points. This study offers data-driven insights to improve BBF predictability, providing valuable guidance for future space weather monitoring and theoretical research.

KEYWORDS

parameter prediction, MultiOutputRegressor, bursty bulk flows, cross-instrument, minimum, maximum, range

1 Introduction

The plasma sheet in Earth's magnetosphere is a highly dynamic region that plays a critical role in transporting energy and particles during geomagnetic active times

(Angelopoulos et al., 1994). Within this region, bursty bulk flows (BBFs)—localized and transient elevation in ion bulk flow speed to the order of hundreds of km/s—are key to understanding how energy is transferred from the magnetotail to the inner magnetosphere (Angelopoulos et al., 1992). Numerous studies have highlighted the critical role of BBFs in the transport of mass, energy, and magnetic flux from the magnetotail to the near-Earth region during geomagnetic activity. Observational analyses, such as those by Nakamura et al. (2001), Nakamura et al. (2002), Nakamura et al. (2005), revealed that BBFs are closely associated with dipolarization fronts and plasma sheet thinning, underscoring their importance in magnetotail reconfiguration. Complementary investigations by Cao et al. (2013) and Yao et al. (2013) further established statistical relationships between BBFs and field-aligned currents or flow bursts, providing insight into their spatial and temporal properties. Forsyth et al. (2008) and Grocott et al. (2004) demonstrated how BBFs influence ionospheric signatures and substorm dynamics, while Henderson et al. (1998) connected BBFs with auroral intensifications. Together, these studies underscore the importance of accurately characterizing BBFs to improve our understanding of magnetospheric dynamics. BBFs are the observational counterpart of plasma-sheet bubbles, which are theoretically defined as depleted magnetic flux tubes containing lower entropy than their neighbors (e.g., Pontius and Wolf, 1990; Birn et al., 2004; Runov et al., 2017). The BBFs or bubbles are often created by magnetic reconnection events (Sitnov et al., 2005; Birn et al., 2011), but may also arise from other explosive magnetotail processes (e.g., Yang et al., 2011; Hu et al., 2011; Sitnov et al., 2019). They can further lead to significant space weather phenomena, such as auroral intensification (Nishimura et al., 2010; Shi et al., 2012) and energetic particle flux enhancements in the inner magnetosphere (Ohtani et al., 2006; Yang et al., 2011).

Although statistical studies that incorporate a set of physical parameters and numerical simulations using advanced MHD or kinetic models have provided invaluable insights, predicting the characteristics of BBFs remains extremely challenging. This difficulty arises from the multiscale nature of magnetotail dynamics, limited observational coverage, and the complex interplay of physical processes driving BBF formation and evolution. Statistical analyses of BBFs involve a number of physical parameters—such as magnetic field, plasma bulk velocity, thermal pressure, temperature and number density—as well as other complex quantities such as magnetic flux transport, specific entropy, electric field and particle distribution functions (e.g., Ohtani, 2004; Liu et al., 2013; Runov et al., 2015; Runov et al., 2017). Like many other statistical approaches, the results often become heavily smoothed, providing only rough estimates of likely ranges. For instance, the left panels of Figure 1 (adapted from Ohtani, 2004) show aggregated measurements that obscure time variations during the BBF injection. Consequently, these statistics cannot yield reliable prediction results for any specific event.

Numerical simulations offer an alternative approach. Certain simulations aim to qualitatively explain the variability of BBFs but face considerable challenges in accurately replicating actual events. For instance, Chen and Wolf (1999) formulated an MHD theory to simulate BBF propagation, treating the moving flux

tube as an infinitely thin filament within a 2D stationary medium in MHD equilibrium. Meanwhile, simulations employing the Rice Convection Model demonstrated an increase in energetic particle flux at geosynchronous orbit due to a BBF's deep injection, generating a dipolarization front via coupling with a force equilibrium solver; however, these simulations omitted inertial effects (Yang et al., 2011). Birn et al. (2011) utilized a 3D one-fluid MHD code to study BBF propagation, observing damped oscillations in the near-Earth region. Yet, their model was confined to a rectangular box encompassing only the nightside region, with perfectly conducting boundaries, and the quantitative accuracy of their idealized simulations hinged on the selection of scaling constants. Other simulations incorporate solar wind conditions as inputs for global MHD codes, and may thereby deliver relatively satisfactory predictions (e.g., Ashour-Abdalla et al., 2011; Merkin et al., 2019). However, these simulations are computationally expensive, and the agreement between model and observation is usually limited to only a few events. An example in the center and right panels of Figure 1 [adapted from Merkin et al. (2019)] illustrates an overall good agreement but reveals discrepancies in the precise timing and magnitudes of key parameters.

Building upon prior research, recent advancements in artificial intelligence (AI) technology and an ever-expanding data pool now offer a more robust foundation for improving prediction (Camporeale, 2019; Bortnik et al., 2018). In this study, our ultimate goal is to provide reliable predictions of key BBF properties—such as mean, maximum, minimum and range—once a BBF occurs. This objective involves two main aspects.

First, we aim to maximize prediction accuracy by utilizing as many relevant input parameters as possible. To this end, we optimized the selection of input parameters by analyzing their distribution and relevance to BBF prediction using Kernel Density Estimation (Chen, 2017). We then employed eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), a powerful machine learning (ML) algorithm capable of modeling complex, nonlinear relationships in high-dimensional data, to develop a robust predictive model for the characteristics of key BBF parameters. We further applied the leave-one-feature-out method to quantitatively assess the contribution of each input, identifying the most dominant factor influencing BBFs in a statistical sense.

Second, we focus on enabling cross-instrument prediction. As satellites age, certain instruments may reach the end of their operational lifespan and cease to provide measurements. In addition, some satellite missions are originally designed to carry only a specific type of payload, resulting in incomplete observational coverage of key space weather events. For instance, the GOES satellites in geosynchronous orbit have accumulated decades of magnetic field data but lack plasma measurements, while the LANL satellites provide long-term plasma data but lack magnetic field observations. These limitations highlight the need for methods that can compensate for missing data. To address this, our study emphasizes cross-instrument prediction—leveraging complementary data from different payloads to estimate unmeasured BBF-related variables. By employing machine learning models trained on both plasma and magnetic field parameters, we can supplement incomplete datasets and improve the utility of existing satellite

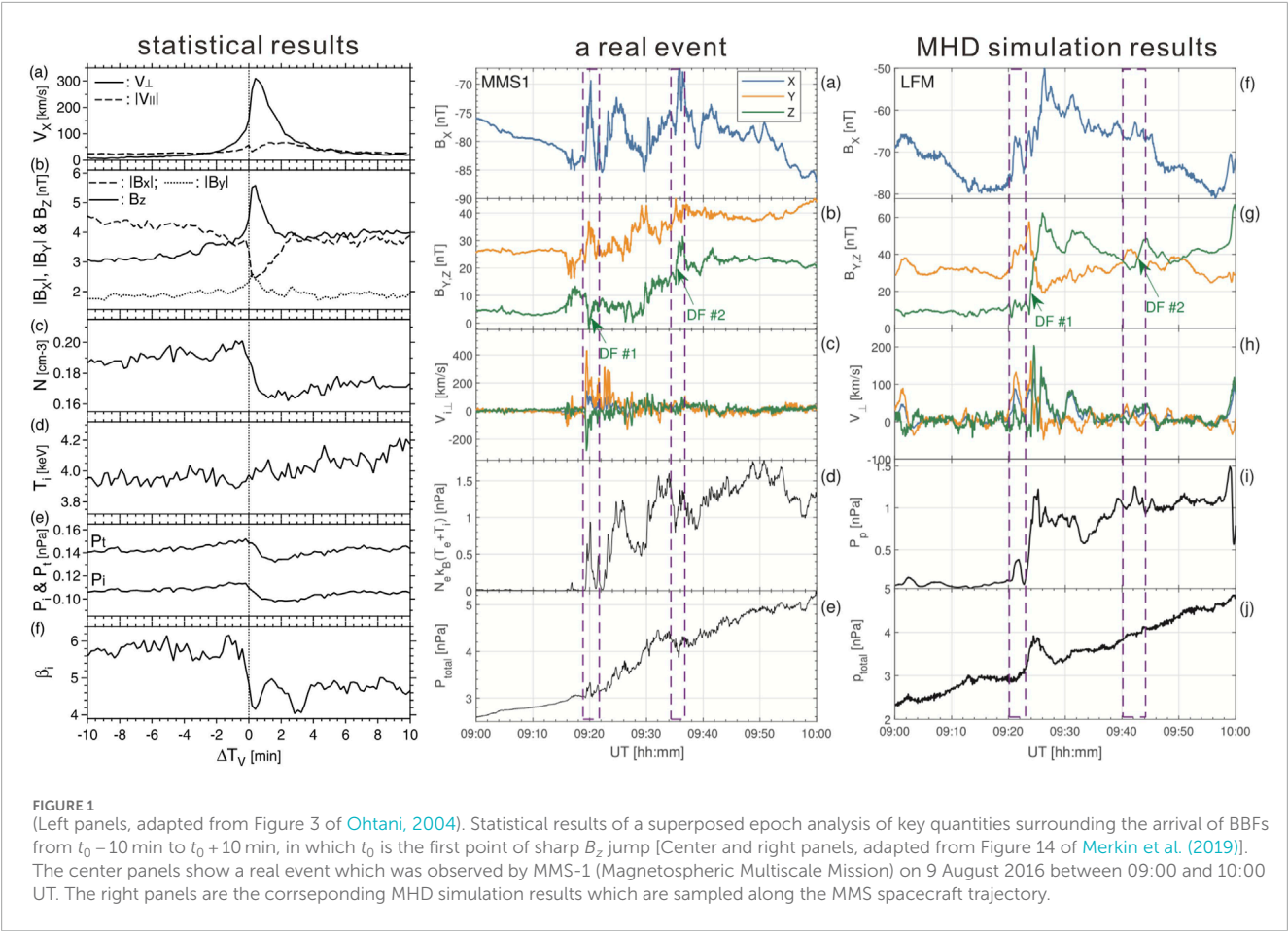


TABLE 1 Summary of machine learning model inputs and targets.

Dataset	Structure	Variables	Features
inputs	BBF	Magnetic parameters: $B_x, B_y, B_z, B , \theta_B, P_m$; Moments and related parameters: $V_{ix}, V_{iy}, V_{iz}, V , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , N_p, N_e, T_p, T_e, P_p, T_i/N_p, T_i/T_e, P_i/N_i^{5/3}, P_e/N_e^{5/3}$; Other parameters: $E_x, E_y, E_z, \beta, \Phi$; Positions: x, y, z .	Mean, Median, Standard Deviation, Minimum, Maximum, Range, 1st Quartile, 3rd Quartile
	background	Magnetic parameters: $B_x, B_y, B_z, B , \theta_B, P_m$; Moments and related parameters: $V_{ix}, V_{iy}, V_{iz}, V , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , N_p, N_e, T_p, T_e, P_p, T_i/N_p, T_i/T_e, P_i/N_i^{5/3}, P_e/N_e^{5/3}$; Other parameters: $E_x, E_y, E_z, \beta, \Phi$; Positions: x, y, z .	Mean
targets		Duration of BBF: Δt_{BBF}	
	BBF	Magnetic parameters: $B_x, B_y, B_z, B , \theta_B, P_m$; Moments and related parameters: $V_{ix}, V_{iy}, V_{iz}, V , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , N_p, N_e, T_p, T_e, P_p, T_i/N_p, T_i/T_e, P_i/N_i^{5/3}, P_e/N_e^{5/3}$; Other parameters: $E_x, E_y, E_z, \beta, \Phi$.	Mean, Minimum, Maximum, Range

observations. This approach not only enhances the completeness of BBF-related information but also contributes to a more accurate understanding of magnetospheric dynamics and supports improved space weather forecasting capabilities.

The paper is organized as follows. Section 2 describes the data collection, the inputs and targets for the ML model. Section 3 explores the application of ML techniques for predicting outcomes based on these parameters. For further analysis, we

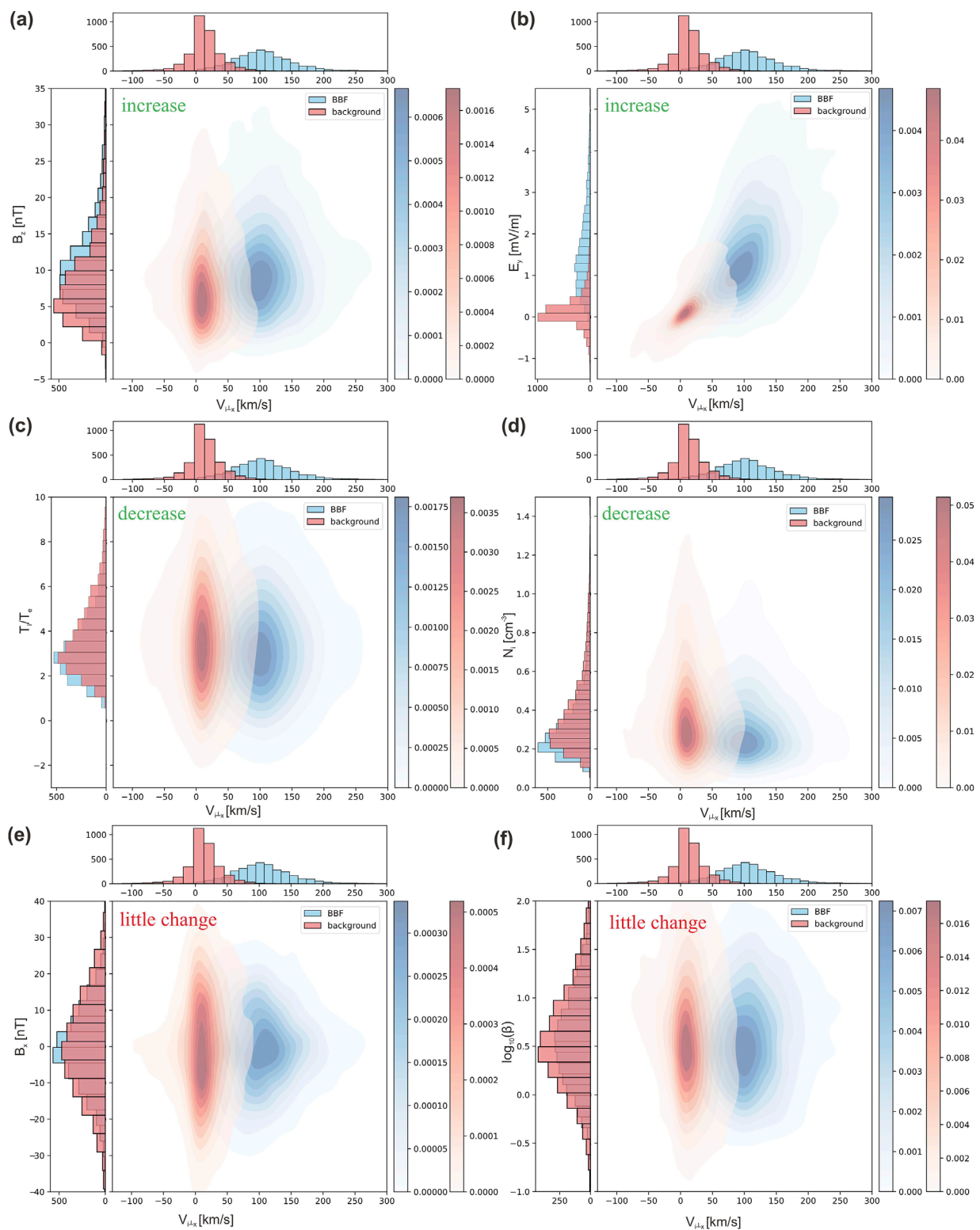


FIGURE 2
Comparison of the probability distributions of the average values of V_{LX} with (a) B_z , (b) E_y , (c) T_i/T_e , (d) N , (e) B_x , and (f) $\log_{10}(\beta)$ between the 3,207 BBF events (blue) and their background environments (red).

TABLE 2 Summary of using optimal combination of parameters to predict the mean, maximum, minimum, and range of BBF parameters.

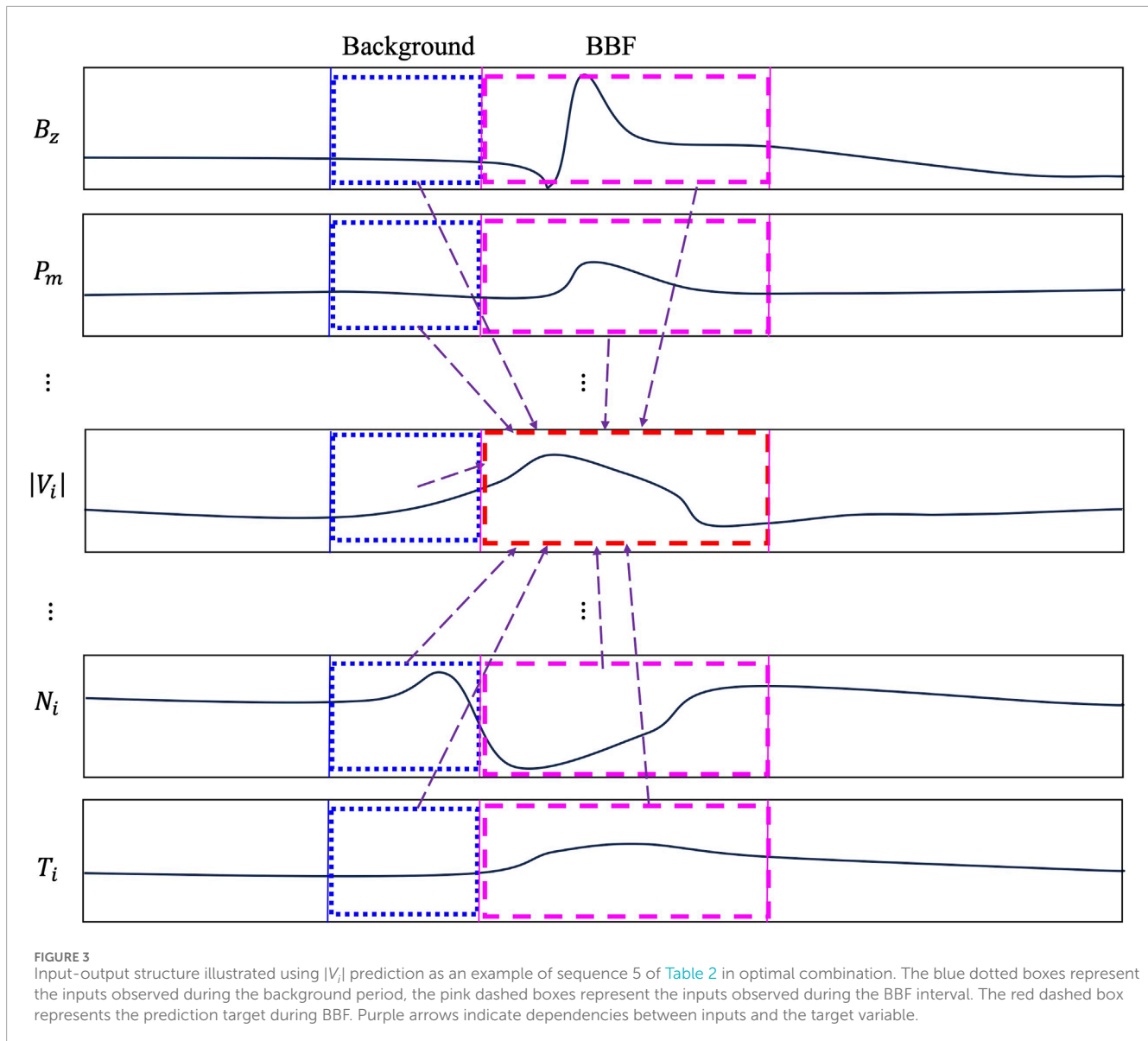
Sequence	Inputs			Targets	
	Variables during BBF and background	Features	Additional variables during background	Variables during BBF	Features
1	$B_x, B_y, B_z, B , \theta_B, P_m, V_{ix}, V_{iy}, V_{iz}, V_i , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , N_i, N_e, T_i, T_e, P_i, T_i/N_i, T_i/T_e, P_i/N_i^{5/3}, P_e/N_e^{5/3}, E_x, E_y, E_z, \beta$ (Φ, x, y, z are excluded.)	BBF interval: Mean, Median, Standard Deviation, Minimum, Maximum, Range, 1st Quartile, 3rd Quartile; Background interval: Mean	-	Δt_{BBF^-}	Mean, Maximum, Minimum, Range
2	$V_{ix}, V_{iy}, V_{iz}, V_i , N_i, N_e, T_i, T_e, P_i, T_i/N_i, T_i/T_e, P_i/N_i^{5/3}, P_e/N_e^{5/3}, x, y, z$ ($B_x, B_y, B_z, B , \theta_B, P_m, V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , E_x, E_y, E_z, \beta, \Phi$ are excluded.)		Mean(B_z)	B_z	
3	$P_i/N_i^{5/3}, P_e/N_e^{5/3}, x, y, z$ ($B_x, B_y, B_z, B , \theta_B, P_m, V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , E_x, E_y, E_z, \beta, \Phi$ are excluded.)		Mean($ B $)	$ B $	
4	$V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , E_x, E_y, E_z, \beta, \Phi$ are excluded.)		Mean(P_m)	P_m	
5	$B_x, B_y, B_z, B , \theta_B, P_m, N_i, N_e, T_i, T_e, P_i, T_i/N_i, T_i/T_e, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta, x, y, z$ ($V_{ix}, V_{iy}, V_{iz}, V_i , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , E_x, E_y, E_z, \Phi$ are excluded.)		Mean($ V_i $)	$ V_i $	
6	$N_i, N_e, T_i, T_e, P_i, T_i/N_i, T_i/T_e, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta, x, y, z$ ($B_x, B_y, B_z, B , \theta_B, P_m, V_{ix}, V_{iy}, V_{iz}, V_i , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , E_x, E_y, E_z, \Phi$ are excluded.)		Mean($ V_{iL} $)	$ V_{iL} $	
7	$B_x, B_y, B_z, B , \theta_B, P_m, V_{ix}, V_{iy}, V_{iz}, V_i , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , T_i, T_e, T_i/T_e, E_x, E_y, E_z, \Phi, x, y, z$ ($N_i, N_e, P_i, T_i/N_i, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta$ are excluded.)		Mean(N_i)	N_i	
8	$N_e, P_i, T_i/N_i, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta$ are excluded.)		Mean(N_e)	N_e	
9	$B_x, B_y, B_z, B , \theta_B, P_m, V_{ix}, V_{iy}, V_{iz}, V_i , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , N_i, N_e, E_x, E_y, E_z, \Phi, x, y, z$ ($T_i, T_e, T_i/T_e, P_i, T_i/N_i, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta$ are excluded.)		Mean(T_i)	T_i	
10	$V_{iLx}, V_{iL} , N_i, N_e, E_x, E_y, E_z, \Phi, x, y, z$ ($T_i, T_e, T_i/T_e, P_i, T_i/N_i, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta$ are excluded.)		Mean(T_e)	T_e	
11	$T_i/T_e, P_i, T_i/N_i, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta$ are excluded.)		Mean(T_i/T_e)	T_i/T_e	
12	$B_x, B_y, B_z, B , \theta_B, P_m, V_{ix}, V_{iy}, V_{iz}, V_i , V_{iLx}, V_{iLy}, V_{iLz}, V_{iL} , E_x, E_y, E_z, \Phi, x, y, z$ ($N_i, N_e, T_i, T_e, T_i/T_e, P_i, T_i/N_i, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta$ are excluded.)		Mean(T_i/N_i)	T_i/N_i	
13	$V_{iLx}, V_{iL} , E_x, E_y, E_z, \Phi, x, y, z$ ($N_i, N_e, T_i, T_e, T_i/T_e, P_i, T_i/N_i, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta$ are excluded.)		Mean(P_i)	P_i	
14	$T_i/T_e, P_i, T_i/N_i, P_i/N_i^{5/3}, P_e/N_e^{5/3}, \beta$ are excluded.)		Mean($P_i/N_i^{5/3}$)	$P_i/N_i^{5/3}$	

designed two primary categories of parameter combinations. Section 4.1 highlights the optimal parameter combination and its corresponding prediction results, while Section 4.2 delves into cross-instrument prediction combination and their associated outcomes. In Section 5, we discuss the challenges and opportunities in predicting the behavior of physical parameters, offering valuable references and ideas for future research.

2 Dataset description

2.1 Observation data and BBF identification

This study utilizes ~17 years of magnetic field and plasma measurements from the five THEMIS probes (Angelopoulos, 2008), covering the period from March 2007 to December



2023. The data include magnetic vectors measured by the Flux Gate Magnetometer (FGM) (Auster et al., 2008), as well as ions (5 eV–25 keV) and electrons (5 eV–30 keV) measured by the electrostatic analyzer (ESA) (McFadden et al., 2008), and ions (25 keV–6 MeV) and electrons (25 eV–1 MeV) measured by the solid-state telescope (SST). The ESA and SST data are combined to provide ion and electron moments such as thermal pressure, density, temperature, and bulk flow velocity (Angelopoulos, 2008). Unless stated otherwise, the Geocentric Solar Magnetospheric (GSM) coordinate system is used. The moments data are interpolated to align with the FGM data due to a timestamp offset, resulting in all parameters having a uniform time resolution of 3 s. Additionally, we compute magnetic field inclination angle θ_B ($\arctan \frac{B_z}{|B_x|}$), the electric field E ($= -V \times B$, assuming frozen-in flux condition), the amount of magnetic flux transported earthward, per unit Y , Φ ($= \int E_y dt$), total plasma thermal pressure P_p (ion pressure plus electron

pressure), magnetic pressure P_m , plasma beta $\beta = P_p/P_m$, specific entropy ($P_i/N_i^{5/3}$ and $P_e/N_e^{5/3}$, where N denotes number density, and the subscriptions i and e denote ions and electrons), as well as the ion-electron temperature ratio T_i/T_e .

Adopting the methodology of Feng and Yang (2023), we identify BBFs using the following traditionally employed criteria: $-20 \leq X \leq -6R_E$, $|Y| \leq 10R_E$, plasma beta $\beta > 0.5$, $B_z > 0 nT$, and $V_{iLX} \geq 200 km/s$ (where V_{iLX} is the X component of the ion bulk velocity perpendicular to the magnetic field). This process yields a total of 3,207 BBF events, which are shown in Supplementary Table S1.

To determine the key features of the BBF parameters such as mean, maximum, minimum, etc., it is essential to know the exact start and end times of each BBF. After conducting experiments, we decide on the following method to establish the start and end times of BBFs, as well as the background periods. Using a three-minute sliding window, this study requires at least one data point within

TABLE 3 MAPE of prediction results using the optimal combination on the validation dataset and test dataset. It should be noted that the result for Δt_{BBF} in the mean column represents Δt_{BBF} itself.

Sequence	Targets	Validation dataset				Test dataset			
		Mean	Max	Min	Range	Mean	Max	Min	Range
1	Δt_{BBF}	33.6%	-	-	-	32.2%	-	-	-
2	B_z	37.8%	17.0%	286.2%	27.2%	26.1%	17.7%	221.4%	30.7%
3	$ B $	14.3%	10.8%	65.0%	22.0%	15.8%	11.5%	80.1%	25.7%
4	P_m	30.1%	22.4%	478.8%	24.8%	34.0%	25.4%	1453.7%	29.9%
5	$ V_i $	19.8%	21.0%	48.4%	27.0%	21.0%	21.9%	32.2%	27.9%
6	$ V_{i\perp} $	18.4%	21.3%	61.8%	26.2%	20.2%	21.9%	47.8%	26.3%
7	N_i	13.1%	14.5%	20.3%	31.0%	12.7%	14.4%	18.8%	31.8%
8	N_e	13.3%	15.4%	26.8%	31.4%	13.1%	14.1%	23.9%	30.4%
9	T_i	10.9%	12.0%	20.4%	34.1%	11.4%	11.9%	18.5%	34.5%
10	T_e	19.9%	25.3%	25.7%	54.1%	19.6%	29.5%	26.4%	51.0%
11	T_i/T_e	17.8%	23.8%	33.6%	44.8%	18.1%	22.0%	30.2%	41.4%
12	T_i/N_i	25.7%	33.0%	34.6%	56.0%	24.5%	32.9%	33.7%	60.7%
13	P_p	12.2%	10.8%	35.0%	23.7%	12.6%	11.6%	36.6%	25.3%
14	$P_i/N_i^{5/3}$	19.1%	24.2%	26.1%	51.9%	20.1%	24.8%	27.0%	53.2%

the sliding window to fulfill the aforementioned criteria of BBFs. The BBF start time is marked by the first instance of $|V_{i\perp}|$ (the ion bulk velocity perpendicular to the magnetic field) exceeding 50 km/s during the sliding process, and the end time is determined when the window continues to later times until finding the first instance of $|V_{i\perp}| \leq 50 \text{ km/s}$. The time duration between its start time and end time is defined as the duration of the BBF, Δt_{BBF} . For the background interval, its start time is defined as the first point 3 minutes prior to the $|V_{i\perp}| \geq 200 \text{ km/s}$. The end time of the background interval coincides with the bubble (BBF) start time.

2.2 Machine learning dataset

From a physics perspective, the properties of BBFs are shaped by both their source conditions and the ambient plasma sheet environment through which they travel. [Sergeev et al. \(2012\)](#) demonstrated that BBFs with comparable reductions in the entropy parameter can penetrate to different locations, depending on the entropy parameter gradient in the background plasma, driven by interchange instability ([Wolf et al., 2009](#)). Comprehensive MHD simulations using parameter-controlled modeling have revealed that both the downstream properties of BBFs in the magnetotail and the background magnetotail configurations can lead to distinct evolution ([Birn et al., 2004](#)). Thus, accurate predictions require incorporating both pre-event background conditions and BBF-specific properties as inputs.

In our machine learning (ML) model, the “input” represents inputs or independent variables, while the “target” represents the outputs or dependent variables. Each variable is characterized by multiple statistical features. As listed in [Table 1](#), we select 31 variables as inputs. For each input variable, we extracted eight features within the BBF interval—mean, maximum, minimum, range (maximum minus minimum), standard deviation, median, first quartile, and third quartile—along with one feature (mean) calculated from the pre-BBF background interval.

The targets include the BBF duration and a subset of physical variables considered predictable, each represented by four features: mean, maximum, minimum, and range during the BBF period. All predictions are conducted within this defined parameter space. To build and evaluate the model, a total of 3,207 BBF events are randomly split into training, validation, and test sets using a 7:1.5:1.5 ratio.

2.3 Selection of predictor variables

To address potential overfitting in the machine learning model and ensure that the predictions reflect meaningful physical connections, we aim to minimize the inclusion of irrelevant variables that could introduce noise or reduce the model’s predictive accuracy. This is achieved by analyzing the distributions of various variables during the background and BBF periods using kernel density estimation (KDE). By comparing the differences between these

TABLE 4 RMSE of prediction results using the optimal combination on the validation dataset and test dataset. It should be noted that the result for Δt_{BBF} in the mean column represents Δt_{BBF} itself.

Sequence	Targets	RMSE units	Validation dataset				Test dataset			
			Mean	Max	Min	Range	Mean	Max	Min	Range
1	Δt_{BBF}	s	75.2	-	-	-	59.1	-	-	-
2	B_z	nT	2.8	4.0	3.8	4.7	2.7	3.9	3.7	4.9
3	$ B $	nT	3.2	3.9	4.1	4.3	3.5	3.8	4.2	4.7
4	P_m	nPa	0.070	0.118	0.064	0.095	0.075	0.103	0.058	0.093
5	$ V_i $	km/s	58.7	115.7	28.9	115.4	60.4	114.6	27.7	115.6
6	$ V_{i\perp} $	km/s	42.6	109.5	14.7	112.5	45.7	104.7	14.7	103.3
7	N_i	cm ⁻³	0.072	0.118	0.066	0.114	0.071	0.115	0.060	0.123
8	N_e	cm ⁻³	0.226	0.168	0.221	0.209	0.122	0.179	0.124	0.219
9	T_i	eV	632.1	921.8	646.3	826.3	665.6	925.7	628.9	834.9
10	T_e	eV	373.7	735.6	303.8	694.4	412.6	857.3	275.5	755.3
11	T_i/T_e	-	2.2	3.8	1.8	3.0	3.5	5.3	3.4	3.3
12	T_i/N_i	keV·cm ³	5.0	9.8	3.5	9.1	4.6	9.2	3.6	9.1
13	P_p	nPa	0.051	0.076	0.056	0.067	0.054	0.097	0.058	0.060
14	$P_i/N_i^{5/3}$	nPa·m ⁵	0.399	0.693	0.325	0.680	0.402	0.710	0.338	0.667

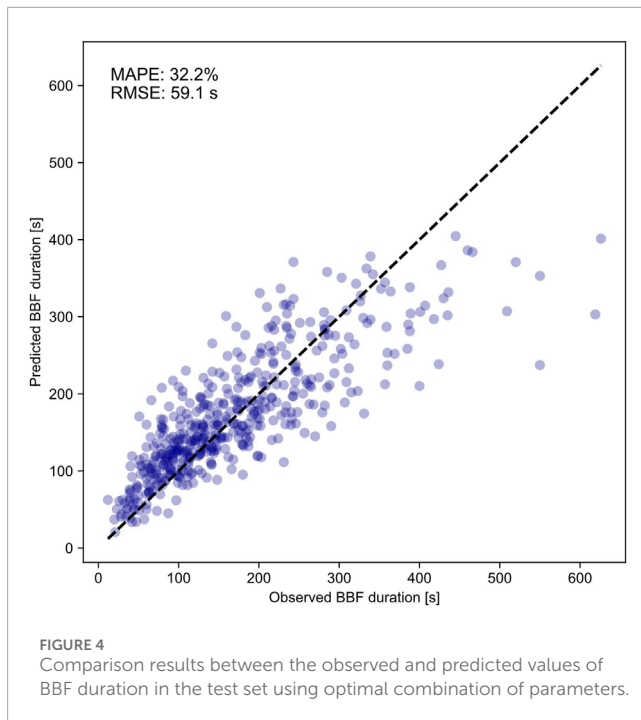
two time periods, we focus on identifying parameters that exhibit significant changes, as these are likely to have the potential for predictivity (i.e. relevant variables). The predictor variables in our model were selected based on their observed variability during the BBF period. Figure 2 illustrates the probability distribution of these parameters, with the central panel showing the overall probability density, and the left and top panels presenting the histogram distributions along each axis. The horizontal axis represents the mean value of $V_{i\perp}$ for all BBF events, which is our most critical BBF velocity criterion. The vertical axis corresponds to the mean value of a specific variable. In the comparison, if the histograms along the vertical axis show that the BBF (blue) and background (red) distributions have a similar shape, and the axes of symmetry of this distribution overlap, we consider there to be no difference. Otherwise, we conclude that the parameter's distribution during the BBF period differs from that of the background. These visualizations provide a comprehensive view of the distributional characteristics and help us isolate variables that deviate notably during the BBF period. For example, panels (a) and (b) confirm increases of B_z and E_y from during the background period to the BBF period; panels (c) and (d) show decreases in T_i/T_e and N_i during BBFs; while panels (e) and (f) indicate no significant difference between the background and BBF in B_x and β . Thus B_x and β can be considered as irrelevant variables.

Using KDE, we statistically analyze the distributions of mean values for the 28 physics parameters (excluding the three positional

parameters) across all 3,207 BBFs during both the BBF and background periods. The analysis confirms that 19 parameters exhibit significant changes during the BBF period compared to the background period. B_z , $|B|$, θ_B , P_m , V_{ix} , $|V_i|$, $V_{i\perp}$, $|V_{i\perp}|$, T_i , T_e , T_i/N_i , $P_i/N_i^{5/3}$, $P_e/N_e^{5/3}$, E_y , and Φ all increase during the BBF period compared to the values during the background periods. In contrast, N_i , N_e , P_p , and T_i/T_e decrease. The remaining 9 parameters show little to no variation and are thus excluded from further analysis. The complete results can be found in the Supplementary Figure S1. These findings highlight the importance of including these 19 parameters that show significant differences.

3 Methodology

Based on the fact that our sample size is low, we choose to use traditional machine learning methods. Traditional machine learning methods involve algorithms that learn patterns from data to make predictions or decisions, and these methods typically require manual feature extraction and selection, where domain expertise is crucial to identify relevant attributes from raw data (Bishop, 2006). Among these models, we evaluate three machine learning models: Support Vector Regression (SVR) (Smola and Schölkopf, 2004), Random Forest (Biau and Scornet, 2016), and XGBoost (Chen and Guestrin, 2016). After assessing their prediction performance on our test dataset, we select the XGBoost model as our prediction method. It builds decision trees sequentially, with each



tree correcting the residuals (errors) of its predecessors. The trees are combined in an additive manner to enhance model performance, and regularization techniques are employed to prevent overfitting (Kakade et al., 2012). Specifically, we utilize a gradient boosting-based regression algorithm called “XGBRegressor” to model non-linear relationships in the data. Since we aim to predict multiple feature values of a specific BBF variable—including the mean, maximum, minimum, and range of the target parameter—we apply the MultiOutputRegressor (Pedregosa et al., 2011) wrapper to manage multiple output variables simultaneously, fitting a separate regressor for each target to ensure flexibility and efficiency.

After establishing the model, the subsequent step is to evaluate and compare its performance in predicting diverse target variables under various parameter combinations. In this study, we adopt a two-pronged approach for performance evaluation. We consider a Mean Absolute Percentage Error (MAPE) (Hyndman and Koehler, 2006) of 35% as the threshold for acceptable prediction. We empirically determined the 35% MAPE threshold through trial and error and manual inspection, balancing predictive accuracy with practical applicability. MAPE is advantageous as it represents errors in percentage terms, offering more intuitive insights compared to metrics like the Root Mean Squared Error (RMSE), which presents results in physical units. However, it should be noted that MAPE has its limitations. It can result in very large errors when the values are small. To provide a more comprehensive assessment, in the following results section, we will present tables for both MAPE and RMSE. This dual - metric presentation allows for a more thorough understanding of the model’s performance, especially considering that our parameters, such as velocity, can range over four orders of magnitude, from 0.1 to 10^3 km/s. By using MAPE as an initial evaluation criterion and supplementing it with RMSE, we aim to facilitate a more complete performance assessment across different variables.

4 Results

In our study, we examine diverse variable combinations and categorize the prediction tasks into two groups. These two groups are “Prediction Using Optimal Combination of Parameters” and “Cross - Instrument Prediction of Magnetic Field Parameters Using Plasma Moments.”

4.1 Prediction using optimal combination of parameters

In the first combination, referred to as the optimal combination of parameters, we utilize as many variables as possible to predict the target feature of a given variable. Our goal is to determine the upper and lower bounds of target variables during the BBF period. The target variable itself and any parameters that can be derived physically must be excluded from the input variables. For instance, when predicting the magnetic field, magnetic pressure cannot be included in the parameters. Through numerous attempts, we also discover that incorporating the mean value of the background of a target variable enhances the accuracy of its prediction, as shown in “Inputs - Additional variables during background” column in Table 2. Additionally, when predicting physical variables, adding positional parameters improves the MAPE value of prediction results by approximately one percentage point. The “Inputs - Variables during BBF and background” column in Table 2 presents all combinations of input physical parameters that are unrelated to the target variables and positional parameters.

This work focuses on predicting the previously selected 19 variables, which would likely change based on the probability distribution analysis. If the MAPE of the mean value is below 35% and at least two of the maximum, minimum, and range MAPE values in the test set are below 35%, we consider the prediction to be valid. This further eliminates six variables, including θ_B , V_{ix} , V_{iLx} , $P_e/N_e^{5/3}$, E_y , and Φ . Ultimately, 13 physical variables are deemed predictable. The BBF duration Δt_{BBF} and 13 other physical parameters constitute all of our target variables, as shown in Table 2. Taking the prediction sequence 5 as an example, the targets are four features, the mean, maximum, minimum and range of $|V_i|$ during the BBF period. The inputs include 21 variables. Among them, twenty are listed in the “Inputs - Variables during BBF and background” column, which does not include velocity. For each of these twenty variables, we calculate eight feature values (mean, median, standard deviation, minimum, maximum, range, 1st quartile, and 3rd quartile) during the BBF interval and one feature value (mean) during the background interval, yielding a total of 180 (9×20) feature values. One additional input is the mean of $|V_i|$ during the background period. Therefore, a total of 181 feature values are used to predict four target feature values for this case.

The data structure used in the model is illustrated in Figure 3. This diagram illustrates how we organize the input and target features of sequence 5 of Table 2 in optimal combination. The input variables include magnetic components B_z , magnetic pressure P_m , ion number density N_i , temperature T_i and other variables that are independent of the velocity. We calculated statistical features of these

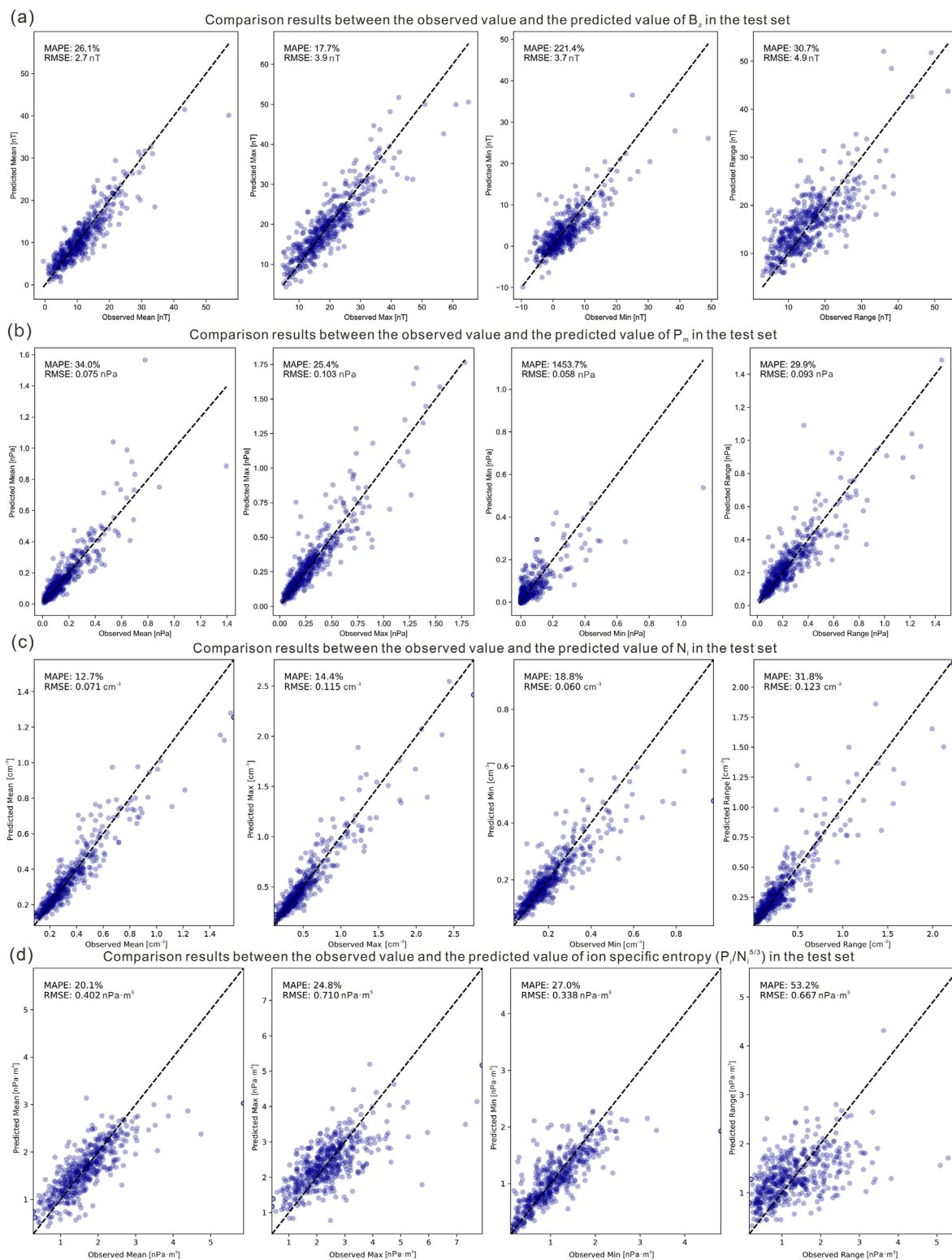
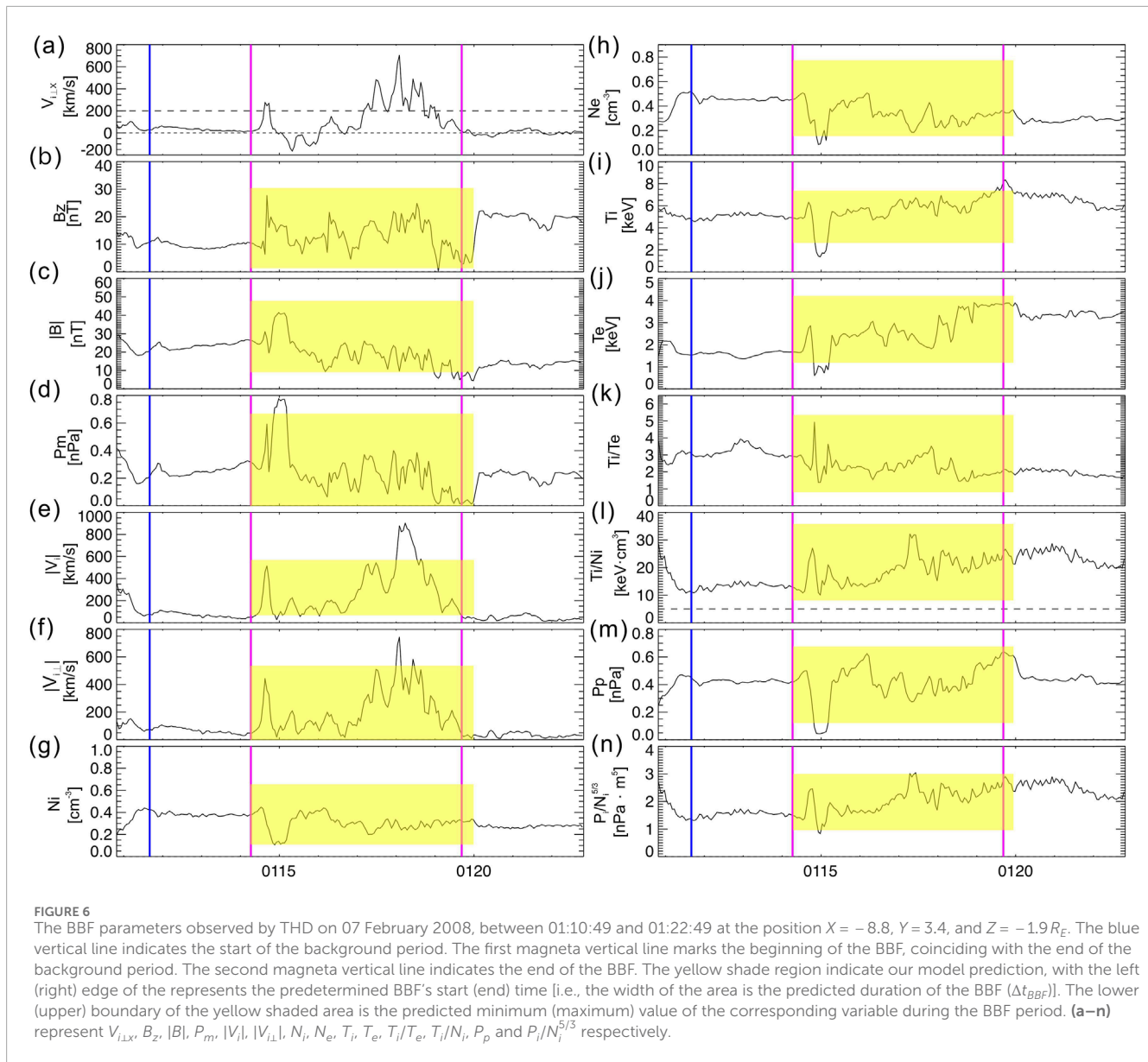


FIGURE 5
Comparison results between the observed values and the predicted values of (a) B_z , (b) P_m , (c) N_i , and (d) $P_i/N_i^{5/3}$ in the test set using optimal combination of parameters.



inputs separately within the background interval (blue boxes) and the BBF interval (pink boxes). These features are then used to predict BBF target characteristics (red box), such as the variation of $|V_{||}|$ during the BBF period.

The statistical metric MAPE of the prediction results is shown in Table 3. The corresponding RMSE results are shown in Table 4.

The comparison between the observed and predicted values for BBF duration in the test set is shown in Figure 4, while the comparisons for B_z , P_m , N_i , and $P_i/N_i^{5/3}$ are shown in Figure 5. Comparisons for other variables are provided in Supplementary Figure S2. Readers may notice that the MAPE for the predicted minimum value of P_m is high in the test dataset, as shown in Figure 5b. This is because the minimum value of P_m is close to zero. To investigate further, we examined the predicted RMSE for the minimum value of P_m , which is 0.058 nPa in the test set as shown in Table 4, confirming that the small value contributes to the large MAPE metric.

Figure 6 illustrates how well the BBF duration and the corresponding maximum and minimum bounds of key physical parameters are predicted using our model. Panel (a) displays the main parameter $V_{||x}$ predicted used to identify BBFs, which helps pinpoint their occurrence. Panels (b)–(n) show the time series of physical parameters measured by the satellite, and the yellow shaded area indicates the predicted BBF duration and the maximum and minimum values of target variables. The results indicate that BBF duration is accurately predicted, differing from actual observations by only about 20 s. The ranges for B_z , $|B|$, T_i , T_e , T_i/T_e , T_i/N_i , P_p , and $P_i/N_i^{5/3}$ are also predicted very well, with the time series during the BBF period largely falling within the shaded area. For P_m , $|V_{||}|$, $|V_{\perp}|$, N_i , and N_e , the predictions for minimum values are more accurate. This is because the maximum value errors are larger due to sharp peaks in velocity's time series and the mean value of the number density in the background is higher than that in the BBF duration. Overall, our parameter combination effectively predicts the variation range of these parameters. If a scientist

TABLE 5 Summary of the most influential input features for different target features.

Target featureduring BBF	The most influential input feature		
	feature name	Δ MAPE (in unit of percentage point)	Δ RMSE
Mean(B_z)	Mean(B_z) duing Background	31.42	2.226 nT
Max(B_z)	Mean(B_z) duing Background	5.91	1.611 nT
Min(B_z)	Mean(N_i) during BBF	300.72	0.045 nT
Range(B_z)	Range(V_y) during BBF	1.36	0.252 nT
Mean($ B $)	Mean($ B $) duing Background	16.30	3.564 nT
Max($ B $)	Mean($ B $) duing Background	6.62	2.677 nT
Min($ B $)	Mean($ B $) duing Background	59.10	2.944 nT
Range($ B $)	Range(P_p) during BBF	1.02	0.184 nT
Mean(P_m)	Mean(P_m) duing Background	53.19	0.101 nPa
Max(P_m)	Mean(P_m) duing Background	22.92	0.100 nPa
Min(P_m)	Mean(P_m) duing Background	1220.15	0.037 nPa
Range(P_m)	Mean(P_m) duing Background	8.78	0.023 nPa
Mean($ V_i $)	Range(P_m) during BBF	1.00	2.298 km/s
Max($ V_i $)	Range(P_p) during BBF	0.68	3.438 km/s
Min($ V_i $)	Max($P_e/N_e^{5/3}$) during BBF	1.87	1.396 km/s
Range($ V_i $)	1st Quartile($P_i/N_i^{5/3}$) during BBF	0.76	−0.735 km/s
Mean($ V_{i\perp} $)	Max(x) during BBF	0.80	1.069 km/s
Max($ V_{i\perp} $)	Mean(β) during Background	0.71	1.761 km/s
Min($ V_{i\perp} $)	Standard Deviation(β) during BBF	6.10	0.327 km/s
Range($ V_{i\perp} $)	Mean($ V_{i\perp} $) during Background	1.13	3.031 km/s
Mean(N_i)	Mean(N_i) duing Background	10.66	0.045 cm ^{−3}
Max(N_i)	Mean(N_i) duing Background	12.32	0.287 cm ^{−3}
Min(N_i)	Mean(N_i) duing Background	7.10	0.022 cm ^{−3}
Range(N_i)	Mean(N_i) duing Background	8.56	0.104 cm ^{−3}
Mean(N_e)	Mean(N_e) duing Background	11.88	0.054 cm ^{−3}
Max(N_e)	Mean(N_e) duing Background	11.17	0.048 cm ^{−3}
Min(N_e)	Mean(N_e) duing Background	10.49	0.032 cm ^{−3}
Range(N_e)	Mean(N_e) duing Background	6.57	0.042 cm ^{−3}
Mean(T_i)	Mean(T_i) duing Background	5.73	392.490 eV
Max(T_i)	Mean(T_i) duing Background	4.38	315.006 eV
Min(T_i)	Mean(T_i) duing Background	6.05	250.835 eV

(Continued on the following page)

TABLE 5 (Continued) Summary of the most influential input features for different target features.

Target featureduring BBF	The most influential input feature		
	feature name	Δ MAPE (in unit of percentage point)	Δ RMSE
Range(T_i)	Mean(T_i) during Background	3.56	65.876 eV
Mean(T_e)	Mean(T_e) during Background	8.47	96.459 eV
Max(T_e)	Mean(T_e) during Background	6.76	67.769 eV
Min(T_e)	Mean(T_e) during Background	16.07	129.465 eV
Range(T_e)	Mean(T_e) during Background	3.42	−0.958 eV
Mean(T_i/T_e)	Mean(T_i/T_e) during Background	8.13	0.267
Max(T_i/T_e)	Mean(T_i/T_e) during Background	8.13	0.267
Min(T_i/T_e)	Mean(T_i/T_e) during Background	4.22	0.235
Range(T_i/T_e)	Mean(T_i/T_e) during Background	10.84	0.338
Mean(T_i/N_i)	Mean(T_i/N_i) during Background	20.38	2.556 keV·cm ³
Max(T_i/N_i)	Mean(T_i/N_i) during Background	12.44	1.963 keV·cm ³
Min(T_i/N_i)	Mean(T_i/N_i) during Background	42.33	2.102 keV·cm ³
Range(T_i/N_i)	Mean(T_i/N_i) during Background	4.33	0.384 keV·cm ³
Mean(P_p)	Mean(P_p) during Background	17.60	0.111 nPa
Max(P_p)	Mean(P_p) during Background	13.39	0.087 nPa
Min(P_p)	Mean(P_p) during Background	22.52	0.057 nPa
Range(P_p)	Mean(P_p) during Background	2.86	0.007 nPa
Mean($P_i/N_i^{5/3}$)	Mean($P_i/N_i^{5/3}$) during Background	14.02	0.217 nPa·m ⁵
Max($P_i/N_i^{5/3}$)	Mean($P_i/N_i^{5/3}$) during Background	7.15	0.174 nPa·m ⁵
Min($P_i/N_i^{5/3}$)	Mean($P_i/N_i^{5/3}$) during Background	22.67	0.185 nPa·m ⁵
Range($P_i/N_i^{5/3}$)	Median(V_{isz}) during BBF	3.13	0.009 nPa·m ⁵

needs to determine the variation range for a specific parameter but encounters measurement or calibration issues with the instruments during that period, they can refer to our parameter combinations for potential solutions. All the results and the corresponding plots for events in the validation and test sets can be found in https://github.com/pinecypressfxd/Project2_MTS_Regression.

After obtaining the prediction results, we further conducted a feature importance analysis to quantitatively assess the contribution of each input in the prediction process. We employed the leave-one-feature-out (LOFO) method, where one feature was removed at a time from the selected input parameter set. The XGBoost model was then retrained using the reduced feature set, and new prediction results on the test set were obtained.

To evaluate the impact of removing each feature, we calculated the differences in prediction performance compared to the baseline

(i.e., predictions using the full input set). Specifically, we defined the performance drop as:

$$\Delta\text{MAPE} = \text{MAPE}_{\text{LOFO}} - \text{MAPE}_{\text{full}}, \Delta\text{RMSE} = \text{RMSE}_{\text{LOFO}} - \text{RMSE}_{\text{full}}$$

where $\text{MAPE}_{\text{LOFO}}$ and $\text{RMSE}_{\text{LOFO}}$ are the prediction errors after removing a single feature, and $\text{MAPE}_{\text{full}}$ and $\text{RMSE}_{\text{full}}$ are the baseline errors.

In most cases, both ΔMAPE and ΔRMSE are positive, indicating that removing the feature degrades prediction performance. However, for a small number of features, we observed negative ΔMAPE or ΔRMSE values, suggesting that excluding those features slightly improved performance, possibly due to noise or redundancy. We ranked all features based on ΔMAPE in descending order to identify the most influential ones in the prediction task.

TABLE 6 Summary of cross-instrument prediction of magnetic field using plasma moments.

Sequence	Inputs		Targets	
	Variables during BBF and background	Features	Variables during BBF	Features
1	Variables Mom: $V_x, V_y, V_z, V_i , N_i, N_e, T_i, T_e, P_i, T_i/N_i, T_i/T_e, P_i/N_i^{5/3}, P_e/N_e^{5/3}, x, y, z (B_x, B_y, B_z, B , \theta_B, P_m, V_{i\perp x}, V_{i\perp y}, V_{i\perp z}, V_{i\perp} , E_x, E_y, E_z, \beta, \Phi$ are excluded.)	BBF interval: Mean, Median, Standard Deviation, Minimum, Maximum, Range, 1st Quartile, 3rd Quartile; Background interval: Mean	B_z	Mean, Maximum, Minimum, Range
2			$ B $	
3			P_m	

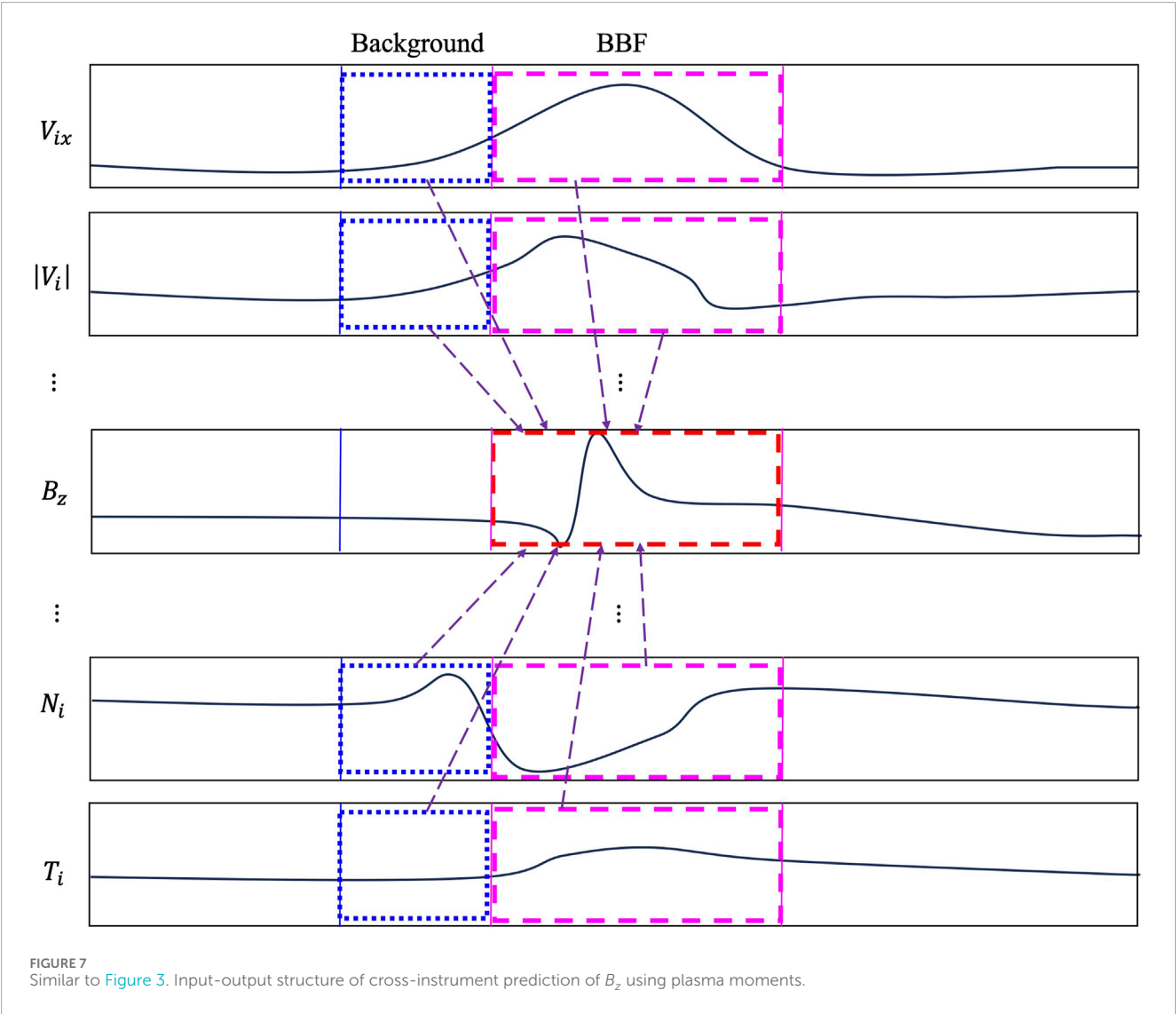


Table 5 summarizes the analysis results. The first column lists each target parameter, while the second identifies the most influential feature for that target. The third and fourth columns report the corresponding Δ MAPE and Δ RMSE results, respectively. Notably, for 42 of the 52 target parameters, the background mean value of the same parameter emerges as the most significant factor, determined by its highest Δ MAPE. This finding underscores

the critical role of background characteristics in predicting target parameter behavior during BBF events. For the remaining 10 targets, features observed during the BBF interval are the most influential inputs. However, no clear physical explanation exists for their primary influencing factors. Numerical simulations solving coupled physics equations could not isolate the effect of a single input while keeping others unchanged in self-consistent modeling. Additionally,

TABLE 7 MAPE of cross-instrument prediction results of magnetic field using plasma moments without additional background inputs.

Sequence	Targets	Validation dataset				Test dataset			
		Mean	Max	Min	Range	Mean	Max	Min	Range
1	B_z	102.3%	22.8%	361.9%	27.3%	42.5%	22.7%	303.5%	30.5%
2	$ B $	29.8%	17.6%	122.1%	22.3%	28.5%	17.8%	135.5%	25.7%
3	P_m	77.9%	41.1%	1414.3%	34.0%	76.9%	42.6%	5517.1%	38.1%

TABLE 8 RMSE of cross-instrument prediction results of magnetic field using plasma moments without additional background inputs.

Sequence	Targets	RMSE units	Validation dataset				Test dataset			
			Mean	Max	Min	Range	Mean	Max	Min	Range
1	B_z	nT	5.1	5.6	5.7	4.7	4.6	5.3	5.3	4.9
2	$ B $	nT	6.7	6.4	7.0	4.4	6.0	5.8	6.4	4.7
3	P_m	nPa	0.140	0.180	0.099	0.126	0.128	0.152	0.103	0.110

we found that removing certain input variables improved prediction accuracy for some target parameters. These variables typically had small impact on the target prediction and were not listed in Table 5 as the most correlated inputs. All detailed results are available in the “feature_selection_result” folder on the referenced website.

4.2 Cross-instrument prediction of magnetic field parameters using plasma moments

We categorize the physical parameters into two main groups: plasma moment parameters and magnetic field parameters. These two categories originate from entirely different payloads. Among these, plasma moments—especially velocity—are critical for identifying BBFs, making their availability a prerequisite for our analysis. Therefore, in our cross-instrument prediction framework, we use plasma moment parameters, along with spacecraft location information, to predict magnetic field measurements.

The parameter combinations we used are summarized in Table 6, this table outlines the structure of input and target variables. The input variables are consistent across all sequences and include plasma moment data, as well as positional coordinates. These variables are statistically characterized by eight features within the BBF interval (mean, median, standard deviation, minimum, maximum, range, first quartile, and third quartile) and one feature (mean) within the background interval. Each sequence targets a different magnetic parameter during the BBF interval.

An example of the data structure for the cross-instrument prediction parameter combination is shown in Figure 7. This figure corresponds to the inputs-outputs structure used for predicting B_z in sequence 1 of Table 6. The input parameters consist of plasma moment variables. Unlike the optimal parameter combination,

the background value of the target variable is excluded from the inputs to enable cross-instrument prediction, even though it has a significant impact on prediction accuracy as we will discuss later.

Through testing, we find that the predictions for B_z and $|B|$ perform well, as shown in Table 7. They have relatively lower MAPE, and among the four features, at least two have a MAPE below 35% in the test dataset. In Table 8, the RMSE values for B_z and $|B|$ are around 4–7 nT, while P_m has an RMSE of approximately 0.10–0.15 nPa. Overall, the prediction accuracy for B_z and $|B|$ is better than for P_m . Figure 8 shows scatter plots comparing the observed and predicted values in the test set.

An example of predicted ranges along with the measurements is shown in Figure 9. The yellow shadow box indicates that we can predict B_z , $|B|$ and P_m with good accuracy. All the results and the corresponding plots for events in the validation, and test sets can be found in https://github.com/pinecypressfxd/Project2_MTS_Regression.

While cross-instrument prediction without additional background data remains challenging, adding observed background mean values as additional inputs can significantly improve performance. In order to illustrate the importance of using accurate background values as ML model inputs in the cross-instrument prediction, we add the mean values of the background periods calculated using actual observations of the background. The MAPE of prediction results are below the 35% MAPE threshold which are shown in Supplementary Table S3, and the time series with predicted bounds for the same event are shown in Supplementary Figure S3. This highlights the importance of background context in achieving robust predictions. The MAPE of these two different additional background values as inputs, are compared in Figure 10. Each bar in the figure represents the average of the three MAPE values (the mean, maximum, and range) on the test set. The error bars represent one standard deviation of

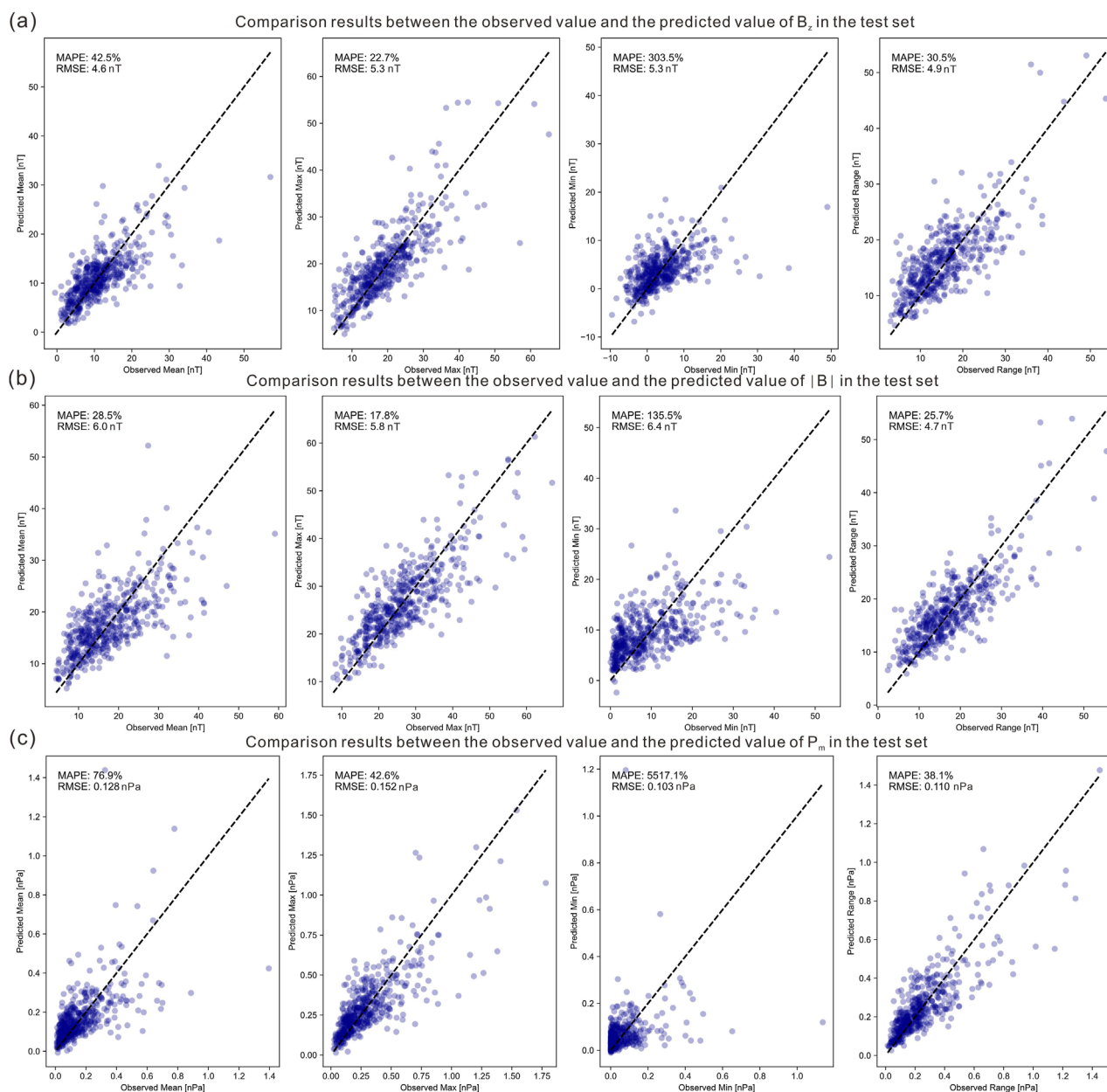


FIGURE 8

Similar to Figure 5. Comparison results between the observed and predicted values of magnetic field targets (a) B_z , (b) $|B|$, and (c) P_m using moment variables combination.

the MAPE of the predictions for these three features. It clearly demonstrates that using the target variable's mean value calculated from actual measurements during the background period can significantly enhance prediction accuracy. For example, for the predicted mean, maximum, and range values during the BBF period in the test dataset, we can see that when using observed background values as additional inputs, the average MAPE results for B_z and $|B|$ decrease by five to twenty percentage points. The reduction for P_m is even more significant.

5 Discussion and summary

Our initial objective was to predict the entire time series of target parameters throughout the BBF period. However, this idea proved impractical due to three primary challenges. First, data limitations significantly constrained our analysis. Despite utilizing all available THEMIS BBF observations, the scarcity of space weather events resulted in only 3,207 BBFs, which was an insufficient sample size for robust time series predictions. Second, the complexity and

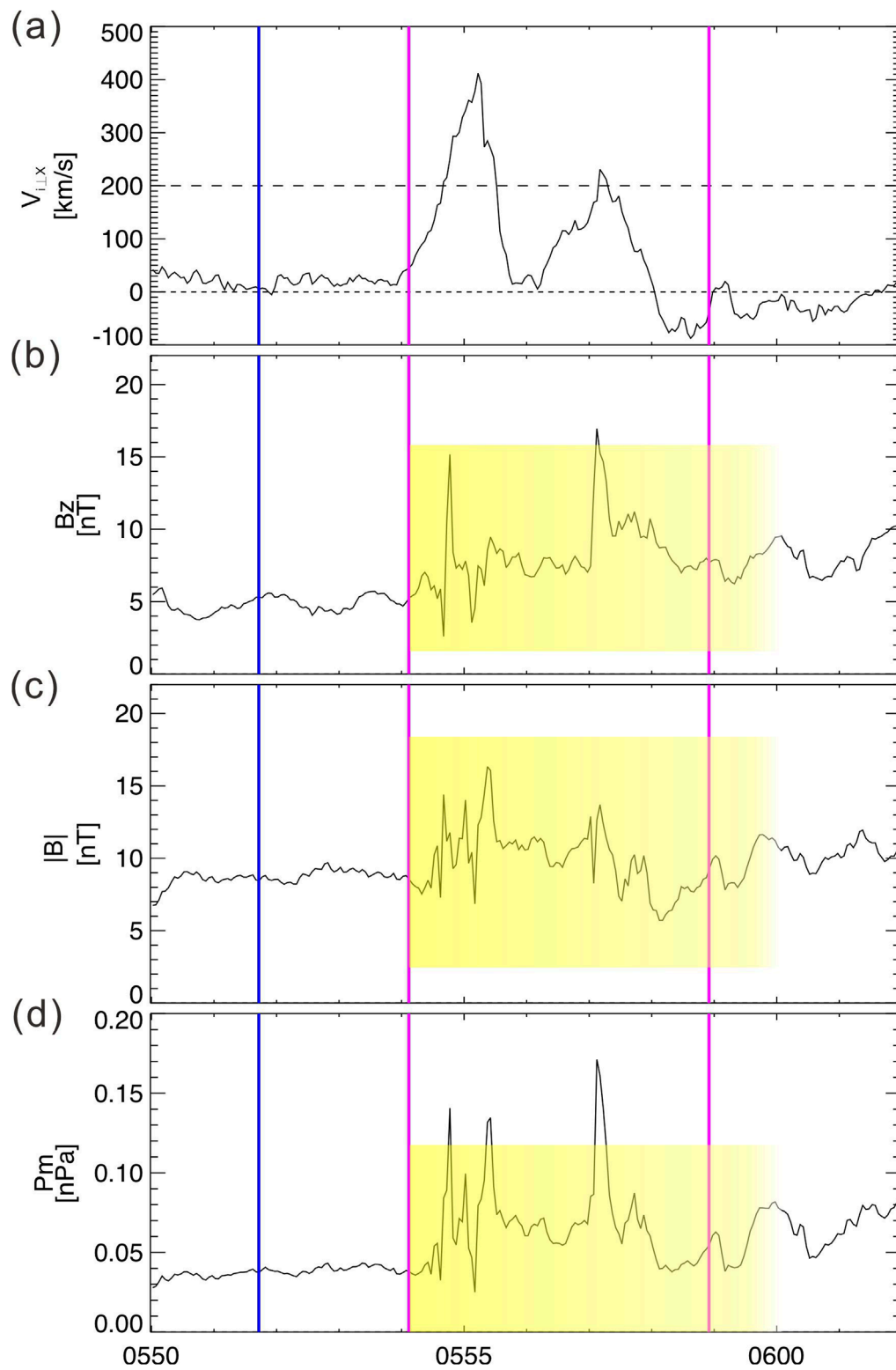


FIGURE 9

Similar to Figure 6. Predicted ranges using cross-instrument prediction of magnetic field measurements using plasma moments without additional background variables as inputs. This event was observed by THE on 07 March 2008, between 05:49:59 and 06:01:59 at the position $X = -10.4$, $Y = 5.1$, and $Z = -1.7R_E$. (a–d) represent $V_{\perp x}$, B_z , $|B|$, and P_m respectively.

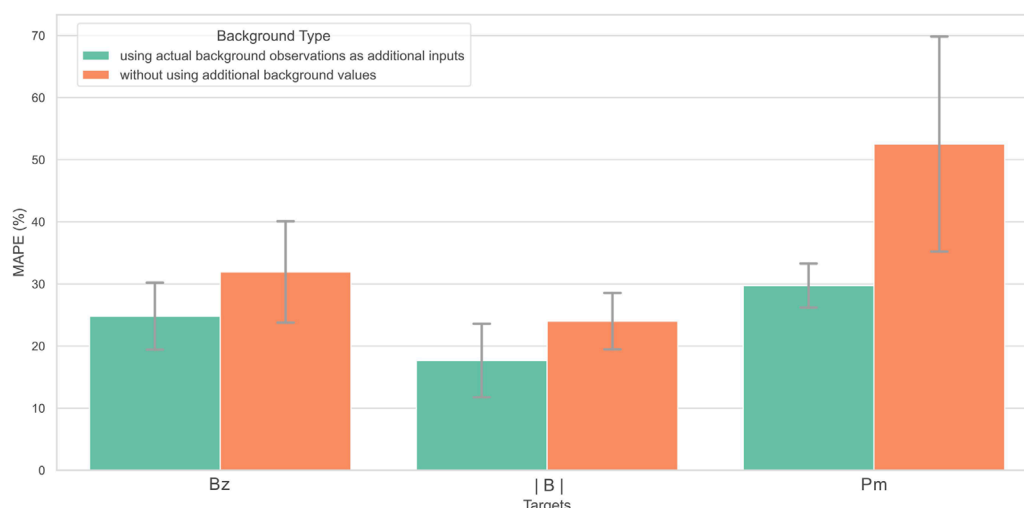


FIGURE 10

Effect of background information on BBF prediction error. The horizontal axis represents the target variables: B_z , $|B|$, and P_m . The vertical axis shows the MAPE of prediction results in the test dataset. The green bars represent predictions using actual background observations as additional inputs. The red bars indicate predictions without using additional background values of B_z , $|B|$, and P_m .

variability of BBF parameter fluctuations posed a major challenge. BBFs often occurred in rapid succession, making their identification difficult even for experienced scientists. This irregularity made it challenging for machine learning models to detect consistent patterns necessary for accurate time-series forecasting. Finally, the presence of excessive microscale variations within BBF time series further complicated prediction efforts. Even after applying strict selection criteria, the inherent variability in parameters—due to background fluctuations and observational noise—remained significant.

To address these challenges, we adopted a feature-based prediction approach, leveraging all available BBF data from 2007 to 2023. Instead of attempting full time-series reconstruction, we focused on identifying predictable features of various parameters. By analyzing the differences between BBF and background periods, we identified parameters that have distinct changing patterns and thus allow for target prediction. We tested different parameter combinations and categorized them into two primary strategies: (1) optimal combinations, where we used as many parameters as possible to maximize prediction accuracy, and (2) cross-instrument prediction, where we leveraged plasma moments parameters to predict magnetic field parameters.

With the optimal combination, we were able to predict BBF duration with a mean absolute percentage error (MAPE) below 35%. Furthermore, we achieved good accuracy for predictions of the maximum and minimum ranges of thirteen key physical parameters, including B_z , $|B|$, P_m , $|V_i|$, $|V_\perp|$, N_i , N_e , T_i , T_e , P_p , T_i/N_i , T_i/T_e , and $P_i/N_i^{5/3}$, with an average prediction error also remaining below 35%. In addition, we conducted a feature importance analysis based on the optimal parameter combinations. The results indicate that the mean values during the background period often play a major role in predicting the corresponding variable's characteristics during the BBF interval. Among the 52 target features, the background mean

serves as the primary influencing factor for prediction accuracy in 42 of them. The physical mechanisms underlying these statistical outcomes present a compelling avenue for future exploration through theoretical analysis or numerical simulations.

For cross-instrument predictions, the results indicate that including accurate target background as an additional input significantly improves the prediction accuracy. By leveraging improved background estimates, we can enhance the reliability of predictions not only for individual BBF parameters but also for broader space weather applications.

Our findings have practical implications for situations where satellites may lack certain payloads or where instrument failures occur. By supplementing missing BBF parameters, our model enhances data utility for space weather research. Expanding the dataset with more comprehensive observations is crucial for improving prediction accuracy. Moving forward, we anticipate that using larger and more diverse datasets (for example, augmenting Geotail, Cluster, and MMS observations) will substantially enhance the model's performance, bringing us closer to a more precise characterization of BBF dynamics.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#), further inquiries can be directed to the corresponding author.

Author contributions

XF: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization,

Writing – original draft, Writing – review and editing. JY: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing. JB: Writing – review and editing, Conceptualization, Formal Analysis, Methodology, Supervision, Validation. C-PW: Formal Analysis, Methodology, Validation, Visualization, Writing – review and editing. JL: Formal Analysis, Methodology, Validation, Visualization, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by Grant 42174197 of the National Natural Science Foundation of China (NSFC), Shenzhen Science and Technology Program (Grant JCYJ20220530113402004).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Angelopoulos, V. (2008). The THEMIS mission. *Space Sci. Rev.* 141 (1–4), 5–34. doi:10.1007/s11214-008-9336-1
- Angelopoulos, V., Baumjohann, W., Kennel, C. F., Coroniti, F. V., Kivelson, M. G., Pellat, R., et al. (1992). Bursty bulk flows in the inner central plasma sheet. *J. Geophys. Res.* 97 (A4), 4027–4039. doi:10.1029/91JA02701
- Angelopoulos, V., Kennel, C. F., Coroniti, F. V., Pellat, R., Kivelson, M. G., Walker, R. J., et al. (1994). Statistical characteristics of bursty bulk flow events. *J. Geophys. Res.* 99 (A11), 21257–21280. doi:10.1029/94JA01263
- Ashour-Abdalla, M., El-Alaoui, M., Goldstein, M. L., Zhou, M., Schriver, D., Richard, R., et al. (2011). Observations and simulations of non-local acceleration of electrons in magnetotail magnetic reconnection events. *Nat. Phys.* 7 (4), 360–365. doi:10.1038/nphys1903
- Auster, H. U., Glassmeier, K. H., Magnes, W., Aydogar, O., Baumjohann, W., Constantinescu, D., et al. (2008). The THEMIS fluxgate magnetometer. *Space Sci. Rev.* 141 (1–4), 235–264. doi:10.1007/s11214-008-9365-9
- Biau, G., and Scornet, E. (2016). A random forest guided tour. *TEST* 25 (2), 197–227. doi:10.1007/s11749-016-0481-7
- Birn, J., Nakamura, R., Panov, E. V., and Hesse, M. (2011). Bursty bulk flows and dipolarization in MHD simulations of magnetotail reconnection: BURSTY FLOWS AND DIPOLARIZATIONS. *J. Geophys. Res. Space Phys.* 116 (A1). doi:10.1029/2010JA016083
- Birn, J., Raeder, J., Wang, Y. L., Wolf, R. A., and Hesse, M. (2004). On the propagation of bubbles in the geomagnetic tail. *Ann. Geophys.* 22 (5), 1773–1786. doi:10.5194/angeo-22-1773-2004
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
- Bortnik, J., Chu, X., Ma, Q., Li, W., Zhang, X., Thorne, R. M., et al. (2018). “Artificial neural networks for determining magnetospheric conditions,” in *Machine learning techniques for space weather* (Elsevier), 279–300. doi:10.1016/B978-0-12-811788-0.00011-1
- Camporeale, E. (2019). The challenge of machine learning in space weather: nowcasting and forecasting. *Space weather*. 17 (8), 1166–1207. doi:10.1029/2018SW002061
- Cao, J., Ma, Y., Parks, G., Reme, H., Dandouras, I., and Zhang, T. (2013). Kinetic analysis of the energy transport of bursty bulk flows in the plasma sheet. *J. Geophys. Res. Space Phys.* 118 (1), 313–320. doi:10.1029/2012JA018351
- Chen, C. X., and Wolf, R. A. (1999). Theory of thin-filament motion in Earth's magnetotail and its application to bursty bulk flows. *J. Geophys. Res. Space Phys.* 104 (A7), 14613–14626. doi:10.1029/1999JA900005
- Chen, T., and Guestrin, C. (2016). “XGBoost: a scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (San Francisco California USA: ACM), 785–794. doi:10.1145/2939672.2939785
- Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostat. & Epidemiol.* 1 (1), 161–187. doi:10.1080/24709360.2017.1396742
- Feng, X., and Yang, J. (2023). Plasma-sheet bubble identification using multivariate time series classification. *J. Geophys. Res. Space Phys.* 128 (10), e2023JA031469. doi:10.1029/2023JA031469
- Forsyth, C., Lester, M., Cowley, S. W. H., Dandouras, I., Fazakerley, A. N., Fear, R. C., et al. (2008). Observed tail current systems associated with bursty bulk flows and auroral streamers during a period of multiple substorms. *Ann. Geophys.* 26 (1), 167–184. doi:10.5194/angeo-26-167-2008
- Grocott, A., Yeoman, T. K., Nakamura, R., Cowley, S. W. H., Frey, H. U., Rème, H., et al. (2004). Multi-instrument observations of the ionospheric counterpart of a bursty bulk flow in the near-Earth plasma sheet. *Ann. Geophys.* 22 (4), 1061–1075. doi:10.5194/angeo-22-1061-2004
- Henderson, M. G., Reeves, G. D., and Murphree, J. S. (1998). Are north-south aligned auroral structures an ionospheric manifestation of bursty bulk flows? *Geophys. Res. Lett.* 25 (19), 3737–3740. doi:10.1029/98GL02692
- Hu, B., Wolf, R. A., Toffoletto, F. R., Yang, J., and Raeder, J. (2011). Consequences of violation of frozen-in-flux: evidence from OpenGGCM simulations: BRIEF REPORT. *J. Geophys. Res. Space Phys.* 116 (A6). doi:10.1029/2011JA016667
- Hyndman, R. J., and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *Int. J. Forecast.* 22 (4), 679–688. doi:10.1016/j.ijforecast.2006.03.001
- Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. (2012). Regularization techniques for learning with. *Matrices* 13 (1), 1865–1890. doi:10.5555/2503308.2343703

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. The author verify and take full responsibility for the use of generative AI in the preparation of this manuscript. Generative AI was used to assist with language editing. The authors reviewed and critically revised the AI-generated content to ensure accuracy, originality, and adherence to academic standards.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fspas.2025.1582607/full#supplementary-material>

- Liu, J., Angelopoulos, V., Runov, A., and Zhou, X.-Z. (2013). On the current sheets surrounding dipolarizing flux bundles in the magnetotail: the case for wedgelets. *J. Geophys. Res. Space Phys.* 118 (5), 2000–2020. doi:10.1002/jgra.50092
- McFadden, J. P., Carlson, C. W., Larson, D., Ludlam, M., Abiad, R., Elliott, B., et al. (2008). The THEMIS ESA plasma instrument and in-flight calibration. *Space Sci. Rev.* 141 (1–4), 277–302. doi:10.1007/s11214-008-9440-2
- Merkin, V. G., Panov, E. V., Sorathia, K. A., and Ukhorskiy, A. Y. (2019). Contribution of bursty bulk flows to the global dipolarization of the magnetotail during an isolated substorm. *J. Geophys. Res. Space Phys.* 124 (11), 8647–8668. doi:10.1029/2019JA026872
- Nakamura, R., Amm, O., Laakso, H., Draper, N. C., Lester, M., Grocott, A., et al. (2005). Localized fast flow disturbance observed in the plasma sheet and in the ionosphere. *Ann. Geophys.* 23 (2), 553–566. doi:10.5194/angeo-23-553-2005
- Nakamura, R., Baumjohann, W., Klecker, B., Bogdanova, Y., Balogh, A., Rème, H., et al. (2002). Motion of the dipolarization front during a flow burst event observed by Cluster. *Geophys. Res. Lett.* 29 (20). doi:10.1029/2002GL015763
- Nakamura, R., Baumjohann, W., Schödel, R., Brittnacher, M., Sergeev, V. A., Kubyshkina, M., et al. (2001). Earthward flow bursts, auroral streamers, and small expansions. *J. Geophys. Res. Space Phys.* 106 (A6), 10791–10802. doi:10.1029/2000JA000306
- Nishimura, Y., Lyons, L., Zou, S., Angelopoulos, V., and Mende, S. (2010). Substorm triggering by new plasma intrusion: THEMIS all-sky imager observations. *J. Geophys. Res. Space Phys.* 115 (A7), 2009JA015166. doi:10.1029/2009JA015166
- Ohtani, S., Singer, H. J., and Mukai, T. (2006). Effects of the fast plasma sheet flow on the geosynchronous magnetic configuration: Geotail and GOES coordinated study. *J. Geophys. Res. Space Phys.* 111 (A1), 2005JA011383. doi:10.1029/2005JA011383
- Ohtani, S.-ichi, Shay, M. A., and Mukai, T. (2004). Temporal structure of the fast convective flow in the plasma sheet: comparison between observations and two-fluid simulations. *J. Geophys. Res.* 109 (A3), A03210. doi:10.1029/2003JA010002
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *Mach. Learn. PYTHON*. doi:10.5555/1953048.2078195
- Pontius, D. H., and Wolf, R. A. (1990). Transient flux tubes in the terrestrial magnetosphere. *Geophys. Res. Lett.* 17 (1), 49–52. doi:10.1029/GL017i001p00049
- Runov, A., Angelopoulos, V., Artemyev, A., Birn, J., Pritchett, P. L., and Zhou, X.-Z. (2017). Characteristics of ion distribution functions in dipolarizing flux bundles: event studies. *J. Geophys. Res. Space Phys.* 122 (6), 5965–5978. doi:10.1002/2017JA024010
- Runov, A., Angelopoulos, V., Gabrielse, C., Liu, J., Turner, D. L., and Zhou, X.-Z. (2015). Average thermodynamic and spectral properties of plasma in and around dipolarizing flux bundles. *J. Geophys. Res. Space Phys.* 120 (6), 4369–4383. doi:10.1002/2015JA021166
- Sergeev, V. A., Chernyaev, I. A., Dubyagin, S. V., Miyashita, Y., Angelopoulos, V., Boakes, P. D., et al. (2012). Energetic particle injections to geostationary orbit: relationship to flow bursts and magnetospheric state. *J. Geophys. Res. Space Phys.* 117 (A10), 2012JA017773. doi:10.1029/2012JA017773
- Shi, Y., Zesta, E., Lyons, L. R., Yang, J., Boudouridis, A., Ge, Y. S., et al. (2012). Two-dimensional ionospheric flow pattern associated with auroral streamers. *J. Geophys. Res. Space Phys.* 117 (A2), 2011JA017110. doi:10.1029/2011JA017110
- Sitnov, M., Birn, J., Ferdousi, B., Gordeev, E., Khotyaintsev, Y., Merkin, V., et al. (2019). Explosive magnetotail activity. *Space Sci. Rev.* 215 (4), 31. doi:10.1007/s11214-019-0599-5
- Sitnov, M. I., Guzdar, P. N., and Swisdak, M. (2005). On the formation of a plasma bubble. *Geophys. Res. Lett.* 32 (16), 2005GL023585. doi:10.1029/2005GL023585
- Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics Comput.* 14 (3), 199–222. doi:10.1023/B:STCO.0000035301.49549.88
- Wolf, R. A., Wan, Y., Xing, X., Zhang, J.-C., and Sazykin, S. (2009). Entropy and plasma sheet transport. *J. Geophys. Res. Space Phys.* 114 (A9), 2009JA014044. doi:10.1029/2009JA014044
- Yang, J., Toffoletto, F. R., Wolf, R. A., and Sazykin, S. (2011). RCM-E simulation of ion acceleration during an idealized plasma sheet bubble injection: BUBBLE INJECTION. *J. Geophys. Res. Space Phys.* 116 (A5). doi:10.1029/2010JA016346
- Yao, Z., Sun, W. J., Fu, S. Y., Pu, Z. Y., Liu, J., Angelopoulos, V., et al. (2013). Current structures associated with dipolarization fronts. *J. Geophys. Res. Space Phys.* 118 (11), 6980–6985. doi:10.1002/2013JA019290