Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research xx (2024) xxx-xxx

ADVANCES IN SPACE RESEARCH
(a COSPAR publication)

www.elsevier.com/locate/asr

Machine Learning Models for Predicting Geomagnetic Storms Across Five Solar Cycles Using Dst Index and Heliospheric Variables

D. Sierra-Porta[a,*], J. D. Petro-Ramos[b], D. J. Ruiz-Morales[b], D. D. Herrera-Acevedo[b], A. F. García-Teheran[b], M. Tarazona Alvarado[c]

[a] *Universidad Tecnológica de Bolívar, UTB. Facultad de Ciencias Básicas, Parque Industrial y Tecnológico Carlos Vélez Pombo Km 1 Vía Turbaco, Cartagena de Indias, 130010, Bolívar, Colombia*
[b] *Universidad Tecnológica de Bolívar, UTB. Facultad de Ingeniería, Parque Industrial y Tecnológico Carlos Vélez Pombo Km 1 Vía Turbaco, Cartagena de Indias, 130010, Bolívar, Colombia*
[c] *Universidad Industrial de Santander. Escuela de Física, Car 27 #9, Bucaramanga, 680001, Santander, Colombia*

Abstract

This study aims to improve the understanding of geomagnetic storms by utilizing machine learning models and analyzing several heliophysical variables, such as the interplanetary magnetic field, proton density, solar wind speed, and proton temperature. Rather than relying on traditional correlation-based methods, we employ advanced machine learning techniques to examine the complex relationships between these factors and geomagnetic storms. Our analysis covers a large dataset spanning six solar cycles, including the current 25th cycle, to provide comprehensive insights into the dynamics of these storms.

Our study highlights the significance of the interplanetary magnetic field as a key predictor of geomagnetic storms, challenging previous beliefs that primarily focused on sunspot activity. By using high-resolution data, we uncover new patterns and provide a more detailed analysis of the factors influencing geomagnetic storms. We emphasize the importance of considering a range of heliophysical variables, such as proton temperature and flow pressure, which offer new insights into the complex dynamics driving these storm events.

The application of machine learning models, particularly Random Forest and Gradient Boosting, demonstrated superior predictive accuracy compared to traditional methods. Our results reveal that the Dst-index MIN, scalar B, and alpha/proton ratio are among the most influential factors, accounting for a significant portion of the prediction model's accuracy. These findings underscore the utility of machine learning in identifying critical drivers of geomagnetic activity and enhancing forecast precision.

Additionally, our research underscores the need for comprehensive models that can accurately predict geomagnetic storms by integrating various data sources. This machine learning approach not only improves predictive accuracy but also enhances our understanding of the underlying mechanisms of space weather. The insights gained from this study have important implications for both scientific research and practical applications, such as improving early warning systems for geomagnetic storms and mitigating their potential impacts on Earth.
© 2024 COSPAR. Published by Elsevier Ltd All rights reserved.

Space Weather ; Machine Learning ; Statistical Modeling ; Geomagnetic Storms ; Data Science

*Corresponding author: orcid=0000-0003-3461-1347;
Email address: dporta@utb.edu.co (D. Sierra-Porta)

https://dx.doi.org/10.1016/j.jasr.xxxx.xx.xxx
0273-1177/© 2024 COSPAR. Published by Elsevier Ltd All rights reserved.

1. Introduction

Geomagnetic storms (GmS) are temporary disturbances of the Earth's magnetic field originating mainly from solar activity. As a transient and dynamic phenomenon, GmS arise from the interaction between the solar wind and the Earth's magnetosphere (Gonzalez et al., 1994; Reyes et al., 2021; Lakhina & Tsurutani, 2016). These disturbances vary in intensity y duración y can have a significant impact on both space and terrestrial environments (Mandea & Chambodut, 2020). Consequences include prolonged interruptions in radio communications (Eid et al., 2022; Love et al., 2023), disruptions to power grids (Taran et al., 2023), damage to satellites and space systems (Abraha et al., 2020), northern and southern lights at lower latitudes, and various technological challenges such as satellite collisions and disruptions in GPS navigation systems (Miteva et al., 2023; Zhang et al., 2020). Additionally, there are potential effects on human health and animal behavior (Sarimov et al., 2023; Kiznys et al., 2020; Hall & Johnsen, 2020).

GmS are typically classified by intensity using indicators like the Disturbance Storm Time (Dst) index, which measures the intensity of the disturbance in the equatorial region, and the AE index for auroral activity. The Kp index, often used to measure global geomagnetic activity, is considered less suitable for severe storm analysis compared to Dst and SYMH indices.

A GmS, in terms of Dst index, is commonly defined as an event where the Dst index drops below a certain threshold, such as -50 nT or -100 nT, as initially described by Sugiura (1960) but now widely accepted (Gonzalez et al., 1994). The Dst index measures the disturbance in the Earth's magnetic field, and its minimum value during a storm serves as a benchmark for classifying the storm's intensity. According to a categorization by Loewe & Prölss (1997), geomagnetic storms are divided into categories based on their intensity. Weak storms have Dst values between -30 nT and -50 nT, indicating relatively minor disturbances. Moderate storms, with Dst values between -50 nT and -100 nT, represent a more significant disruption. Strong storms, marked by Dst values between -100 nT and -200 nT, indicate a considerable impact on the Earth's magnetosphere. Severe storms have Dst values ranging from -200 nT to -350 nT, showing substantial disturbances, while great storms, characterized by Dst values below -350 nT, represent the most extreme geomagnetic events.

There is no official classification based on duration alone, but general categories are often used: (a) brief storms (< 6 hours), typically less intense; (b) moderate storms (6-24 hours); (c) prolonged storms (> 24 hours), such as the Halloween Storm of 2003 (Lopez et al., 2004; Hady, 2009); and (d) extreme storms, rare historic events such as the Carrington event of 1859 (Tsurutani et al., 2003; Siscoe et al., 2006).

Previous studies have examined GmS occurrences through various approaches. In a first study Tsurutani et al. (1995), delves into the interplanetary origins of geomagnetic activity, highlighting the significance of high-speed solar wind streams and the interplanetary magnetic field (IMF) during the declining phase of solar cycles. The study reinforces the critical influence of these factors on geomagnetic storms. Also, Abe et al. (2023) analyzed GmS occurrences using Dst and Sunspot Number (SSN) data during solar cycles 20-24, showing that GmS occurrence rates are higher during descending phases. It primarily employs statistical methods to analyze these trends and identifies similar key drivers, such as coronal holes and solar wind streams. Furthermore, Hajra et al. (2021) highlights long-term variations in geomagnetic activity, noting that strong solar cycles tend to exhibit more frequent and intense geomagnetic storms compared to weak cycles. The authors emphasize the role of high-speed solar wind streams from coronal holes, particularly during the declining phases of solar cycles, in driving geomagnetic activity.

Collectively, these studies underscore the importance of heliospheric conditions, particularly during the declining phases of solar cycles, in influencing geomagnetic storm activity. They consistently highlight the role of high-speed solar wind streams and the IMF as significant contributors to geomagnetic phenomena. Additionally, it is revealed that severe and extreme GmS (Dst < -250 nT) seldom occur during low solar activity but rather during periods of very high solar activity and are mostly associated with coronal mass ejections (CMEs) when they occur. It has also been revealed that all high-intensity GmS (strong, severe, and extreme) are mostly associated with CMEs (Echer et al., 2008; Gonzalez et al., 2011; Echer et al., 2013). The results have shown that CMEs are the primary cause of GmS in the ascending, maximum, and descending phases of cycles 23 and 24, followed by CMEs and High-Speed Solar Wind.

Other studies have used correlational analyses to investigate solar and interplanetary factors influencing GmS (Le et al., 2013; Miteva et al., 2023; Rathore et al., 2012; Samwel & Miteva, 2023; Singh Chauhan et al., 2010; Yacouba et al., 2022), with many focusing on specific predictors such as sunspots (Abe et al., 2023; Reyes et al., 2021) and CMEs (Srivastava & Venkatakrishnan, 2002; Nitta et al., 2021).

In this study, we aim to contribute this field in several ways. First, by using monthly resolution data to understand more broadly the conditions that lead to geomagnetic storms. Second, by incorporating the interplanetary magnetic field and other heliophysical variables like proton density, solar wind speed, and temperature, flow pressure and interplanetary magnetic field, to expand the scope of previous research. Third, by using long-term data spanning six solar cycles, including the ongoing 25th cycle, to analyze trends and variability across multiple solar cycles.

Our approach involves robust statistical models, including multiple linear regressions and machine learning models, to capture the non-linear dynamics between the number of GmS and predictor variables. This offers a more comprehensive understanding of the factors influencing GmS and enhances predictive accuracy beyond traditional correlational analyses.

2. Data used in this study: predictors for geomagnetic storms

The data used in this study are from the OMNI2 dataset, which is available in the `https://omniweb.gsfc.nasa.`

gov/ directory of the NASA OMNIWEB website. These data comprise hourly mean values of the interplanetary magnetic field (IMF), solar wind plasma parameters, and various geomagnetic and solar activity indices, as well as energetic proton fluxes.

OMNI2 was developed at the NSSDC (National Space Science Data and Services Center) in 2003 as an evolution of the OMNI data set, initially created in the mid-1970s. These data are collected from various NASA space missions, including: IMP 1, 3, 4, 5, 6, 7, 8 (Fairfield et al., 1981; Paularena & King, 1999), these space probes, also known as Explorers, have contributed significantly to the collection of data on the interplanetary environment near Earth orbit; WIND (Ogilvie & Desch, 1997; Wilson III et al., 2021), is a space mission equipped with a magnetometer, has provided detailed measurements of the solar wind and the interplanetary magnetic field; ACE (Advanced Composition Explorer) (Garrard et al., 1998; Chiu et al., 1998) which is a space probe that has collected precise measurements of solar wind and energetic particles from its orbit around the L1 Lagrange point; and Geotail (Frank, 1994; Schmidt et al., 1995), a joint mission between JAXA (Japan Aerospace Exploration Agency) and NASA; among others.

Variables selected for this study include year; decimal day; time; the hourly average of the magnitude of the IMF, expressed in nanoteslas (nT); proton temperature (PT) and density (PD) in the solar wind, measured in degrees Kelvin and protons/cm^3; plasma wind speed (PS) of the solar wind plasma, measured in kilometers per second (km/s); Alpha/Proton ratio (A/P), the ratio of alpha particles to protons in the solar wind, flow pressure (FP) represents the density of protons in the solar wind, measured in nanopascals (nPa), geomagnetic index Kp, sunspot number (R) which is a measure of the number of sunspots present on the Sun at a given time, Dst index provides a measure of the intensity of the Earth's magnetic field in the magnetic tail region during GmS; and F10.7 index for solar activity at wavelength 10.7 centimeters, expressed in solar flux units (sfu).

We use the Dst geomagnetic index, which measures the Earth's ring current and is expressed in nanoteslas (nT). The Dst index is a global measure of geomagnetic storm intensity and is calculated from magnetic deviations recorded at various magnetic stations near the equator. Negative values of the Dst index indicate a greater disturbance in the magnetosphere, and the lower the value, the more intense the storm. It is commonly used to determine the number and intensity of geomagnetic storms, as it is one of the primary indicators for this purpose.

In this study, it is important to note that the Sunspot Number is recorded with daily resolution, since we are interested in correlating this number with the occurrence of GmS, we need to homogenize the resolution of the data so that they are all on the same time scale. Therefore, we perform a resampling of the data to group them into day intervals, which allows us to consistently compare the sunspot number with other variables measured at an hourly resolution.

In addition, this study makes use of data collected over six solar cycles, starting from 1964 (solar cycle 20) to the present (solar cycle 25, which is still in an rising phase). This broad time window allows us to examine long-term trends in solar activity and their relationship to the occurrence of GmS. By spanning multiple solar cycles, we can get a more complete picture of how solar activity varies and how this affects the incidence of storms over time.

Finally, the number of GmS (Total ST) per day is defined as the number of records where the Dst index falls below -50 nT, a threshold that signifies the presence of a geomagnetic storm. This threshold allows for the capture of all storms, from moderate to severe, and provides a comprehensive analysis of geomagnetic activity during the study period. By using this criterion, we can effectively capture and analyze the frequency and intensity of GmS based on the Dst index (Borovsky & Shprits, 2017; Loewe & Prölss, 1997).

The temporal distribution of geomagnetic storms according to the Dst index across solar cycles 20 to 25 reveals several patterns in storm frequency is shown in Figure 1. The data shows significant variability between cycles. These differences highlight the unique characteristics of each solar cycle and the importance of considering long-term trends when analyzing geomagnetic activity.

In addition to considering the Dst index to obtain the number of geomagnetic storms, we are also including the following variables to develop machine learning regression models that serve as predictor variables: sunspot number (R), solar radio flux at 10.7 cm (F10.7), proton temperature, proton density, plasma speed, alpha/proton ratio, flow pressure, and field magnitude average (|B|). These variables have been selected based on their relevance and potential influence on geomagnetic storm activity as indicated by prior research.

The inclusion of these additional heliophysical variables allows for a more comprehensive analysis, capturing the complex and nonlinear interactions within the solar-terrestrial environment. By leveraging these diverse datasets, we aim to improve the predictive power and accuracy of our models, ultimately providing deeper insights into the mechanisms driving geomagnetic storms.

3. Methods and techniques

This section details the methodology used for the implementation of the regression models and the selection of the best models. Five different regression models were employed: Multiple Linear Regression (MLR), Random Forest (RF), Gradient Boosting (GB), AdaBoost (AB), and ExtraTrees Regressor (ET). All implementations were carried out using the Python 3.10 library ecosystem, specifically the SKLEARN module, in a development environment running on an Intel Core i7 8-core processor computer.

The objective of this approach is to develop more robust regression models to understand the relationships between the number of geomagnetic storms and certain solar dynamics parameters, and to identify which of these parameters most significantly influence the modeling of geomagnetic storms.

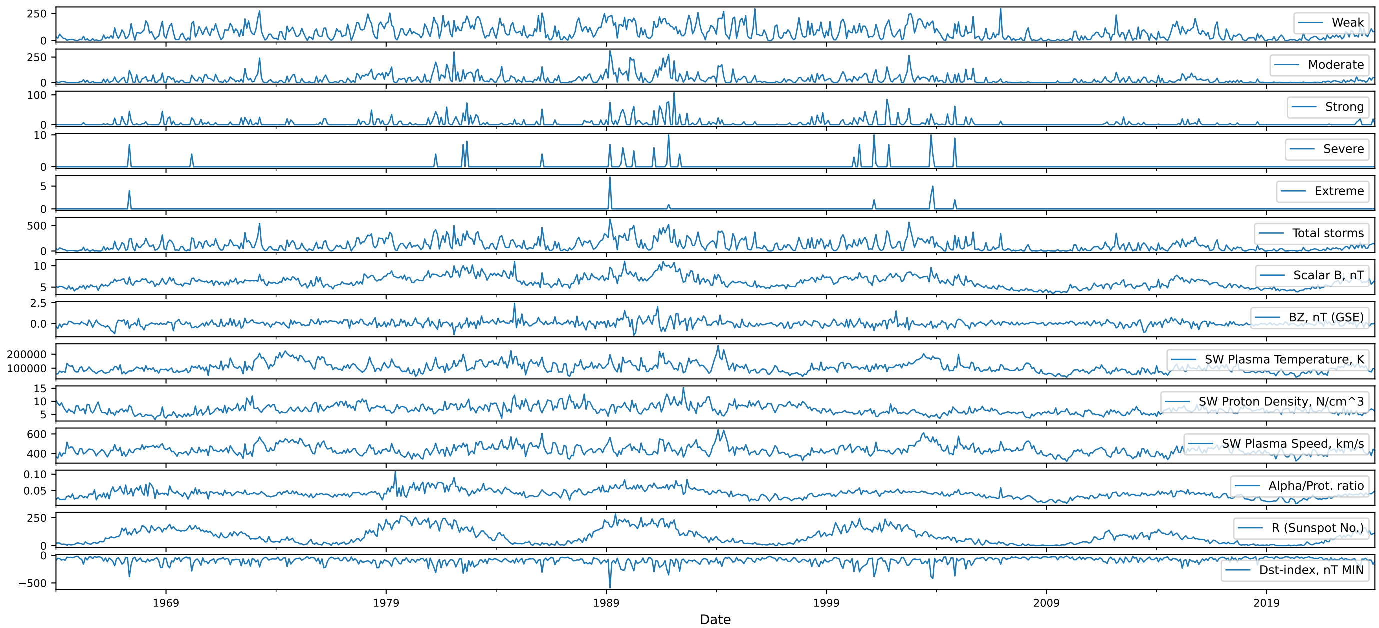

Fig. 1. Temporal distribution and evolution of geomagnetic storms according to Dst index, for cycles 20 to 25 (still in progress) from the data collected and processed. Additionally, the temporal distribution of the heliopheric dynamics supporting variables is shown. The figure shows a clear correlative trend between the number of geomagnetic storms and some solar indices, particularly sunspots and interplanetary magnetic field.

3.1. Data and model preparation

In the preparation of the data, we chose not to normalize or standardize the data. Instead, the models were constructed using the typical magnitudes and values of each variable. This decision was made considering that the characteristics of the predictor variables presented scales and ranges of values inherent to their respective domains, which allowed the models to more accurately capture the relationships and variations within the original data. Consequently, the introduction of artificial biases or distortion of intrinsic relationships between variables was avoided by maintaining the natural scale and distribution of the data during the modeling process.

Prior to the implementation of the regression models, the data were properly prepared. A partitioning of the data set into training and test was applied, using a test length ratio of 70:30 (30% test size from whole dataset). This partitioning was performed randomly to ensure representativeness of both sets.

Each of the regression models was implemented using the sklearn library. The parameters used for the selection of the best model in each case are described below.

For the Multiple Linear Regression model, the implementation was carried out using the default settings of the sklearn library (Pedregosa et al., 2011). For the machine learning models Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, hyperparameter optimization was conducted using TPOT (Olson & Moore, 2016; Moore et al., 2023), which automates the machine learning pipeline design by leveraging genetic programming. The hyperparameter tuning (Adnan et al., 2022; Alibrahim & Ludwig, 2021) with Cross-Validation strategy (Schaffer, 1993; Nti et al., 2021) explored various settings for the number of trees (N_ESTIMATORS), learning rate (LEARNING_RATE), maximum tree depth (MAX_DEPTH), and other relevant parameters specific to each model. The parameters tested for the Random Forest and Gradient Boosting models included N_ESTIMATORS = [100, 200, 300, 500, 700, 1000, 1200, 1400, 1600, 1800, 2000], MAX_FEATURES = ['auto', 'sqrt', 'log2'], MAX_DEPTH = [50, 100, 150, 200, 500], MIN_SAMPLES_SPLIT = [2, 5, 10, 14, 16, 18], MIN_SAMPLES_LEAF = [1, 2, 4, 6, 8, 12, 16, 20], CRITERION = ['absolute_error', 'friedman_mse', 'squared_error', 'poisson'], and CCP_ALPHA = [0.0, 0.01, 0.1, 0.001, 0.0001]. For the Gradient Boosting Regressor, additional parameters such as LOSS = ['squared_error', 'absolute_error', 'huber', 'quantile'] and LEARNING_RATE = [0.001, 0.005, 0.01, 0.05, 0.1, 0.5] were evaluated. TPOT's genetic algorithm was configured with parameters such as generations (=6), population size (=35), offspring size (=30), and an early stopping criterion (=12) to optimize model selection and parameter tuning.

This approach allowed for an efficient exploration of the hyperparameter space, ultimately identifying the best-performing models and configurations for each algorithm. The performance of the optimized models was evaluated on an independent test set, ensuring robust generalization and reliability of the predictive capabilities.

To calculate the number of geomagnetic storms, we resample the Dst variable from its original minute resolution by counting all events within 30-day intervals where Dst is less than -50 nT. This approach aggregates the data into a more manageable form while capturing the frequency of significant geomagnetic storm events over time. During these periods, we also compute the average values for other variables of interest, including sunspot number, solar radio flux, proton temperature, proton density, plasma speed, alpha/proton ratio, flow pressure, and field magnitude average ($|B|$).

Averaging these variables over intervals helps to smooth out short-term fluctuations and reduces the impact of outliers and noise, which is important given the high variability of heliophysical data. This strategy provides a clearer view of the relationships between variables and enhances the robustness of our regression models, allowing for more reliable predictions of geomagnetic storm activity.

At the end of this process, we obtain a dataset that contains information for each variable measured within the same time intervals. This ensures that all variables correspond accurately to the same 15-day periods, providing a consistent temporal framework for analysis. By aligning the data in this manner, we can directly compare the different variables and their influence on geomagnetic storm activity within the same time frames. This harmonized dataset facilitates the development of regression models by ensuring that each observation includes a comprehensive set of predictor variables, all measured over identical intervals. Additionally, this approach allows for more straightforward integration and comparison of different datasets, which might have varying original resolutions, by standardizing them to a common temporal scale. The resulting dataset is thus not only comprehensive but also tailored to maximize the analytical robustness of our machine learning models, enhancing our ability to identify and understand the relationships between heliospheric conditions and geomagnetic storms.

3.2. Machine Learning models

In the study, five regression models were implemented, each with its own characteristics and philosophy of use. Starting with MLR, this model is an extension of simple linear regression and seeks to establish a linear relationship between a dependent variable and multiple independent variables. It is a classic model that assumes a linear relationship between the variables and is useful when seeking to understand the relative contribution of each predictor variable to the target variable. Although it is simple and easy to interpret, it may not capture nonlinear relationships between variables.

In contrast, RF (Breiman, 2001; Hastie et al., 2009; Biau & Scornet, 2016) is a more complex model that combines multiple decision trees to make predictions. Each tree is independently trained on a random sample of the data and produces a prediction. By averaging the predictions of all the trees, RF reduces variance and overfitting, making it robust to noisy data or data with high dimensionality. In addition, RF is capable of handling missing data and categorical variables, making it a versatile and powerful option for regression.

Another ensemble is GB (Natekin & Knoll, 2013; Bentéjac et al., 2021; Friedman, 2002) model that combines multiple decision trees, but unlike RF, the trees are added sequentially, gradually improving the accuracy of the model. At each iteration, GB adjusts a new tree to correct the prediction errors of the existing model. This boosting technique allows building a highly predictive model, especially suitable for regression problems with high dimensionality data. However, GB can be more prone to overfitting than RF and may require careful parameter tuning.

On the other hand, AB (Ying et al., 2013; Schapire, 2003; Drucker, 1997) is a boosting algorithm that combines multiple weak classifiers to form a strong classifier. Unlike GB, which focuses on reducing model bias, AB focuses on reducing variance by giving more weight to misclassified instances at each iteration. AB is robust to overfitting and can handle unbalanced or noisy data, making it suitable for regression problems with complex data sets.

Finally, the ET (Geurts et al., 2006) is a variant of the Random Forest algorithm that is characterized by making random decisions during the construction of each decision tree. This randomness can lead to greater diversity among trees and, in some cases, better predictive performance than RF. ET is particularly useful when seeking to reduce overfitting and increase model stability in small or noisy data sets.

The hyperparameter optimization and training times for the machine learning models varied significantly depending on the complexity of each algorithm and the size of the hyperparameter search space. Using an Intel Core i7 processor with 16 GB of RAM running Ubuntu, the training and hyperparameter tuning times were recorded as follows: Random Forest (RF) took approximately 27.56 minutes (1653.61 seconds), Gradient Boosting (GB) required approximately 85.17 minutes (5110.21 seconds), HistGradientBoosting (HGB) completed in approximately 4.08 minutes (244.94 seconds), Extra Trees (ET) finished in approximately 2.12 minutes (127.75 seconds), and AdaBoost (AB) required approximately 24.07 minutes (1444.15 seconds). The total execution time for all models was approximately 60 minutes.

These differences in training time can be attributed to the inherent computational complexity of each model and the extent of the hyperparameter space explored. For instance, AdaBoost and Gradient Boosting involve sequential training processes and often require more time to converge to an optimal solution, whereas algorithms like Extra Trees benefit from parallel training processes, resulting in faster execution. The use of TPOT's automated pipeline optimization and its genetic programming approach also contributed to the differences in training times, as TPOT dynamically explores various model architectures and hyperparameter configurations to identify the most effective solutions.

This trade-off between model accuracy and computational efficiency is a critical consideration when deploying machine learning models in production environments, particularly for applications requiring real-time or near-real-time predictions. The computational resources and time required for training must be balanced with the expected performance improvements achieved through hyperparameter tuning.

3.3. Evaluation techniques for models

To select the best model describing the number of geomagnetic storms as a function of the predictor variables, several evaluation metrics will be used to compare the performance of the models in a comprehensive manner. These metrics include mean absolute percentage error (MAPE), r2-score, root mean squared error (RMSE), Akaike information criterion (AIC), Bayesian Information Criterion (BIC) and the correlation coefficient (CC) between the actual values (data) and those predicted by the models.

The MAPE is an error measure that calculates the average percentage error between the actual values and the values predicted by the model. It is calculated as the average of the absolute value of the difference between the actual values and the predicted values, divided by the actual value, and multiplied by 100,

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{1}$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value. Their difference is divided by the actual value. The absolute value of this ratio is summed for every predicted point in time and divided by the number of fitted points n.

The r2-score, also known as the coefficient of determination, is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as the proportion of the variance explained by the model to the total variance in the data. In other words, The RMSE is a measure of the root mean squared error between the actual values and the values predicted by the model. This metric provides a measure of the accuracy of the model predictions in terms of the scale of the original data.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}. \tag{2}$$

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are measures used to compare statistical models based on their fit and complexity. These measures penalize more complex models, favoring those that achieve a good fit with a smaller number of parameters.

Finally, the correlation coefficient between the actual values and those predicted by the models provides a measure of the linear relationship between these two variables. A correlation coefficient close to 1 indicates a strong positive correlation between model predictions and actual values, while a coefficient close to -1 indicates a strong negative correlation. A coefficient close to 0 indicates a weak or no correlation between the variables. This coefficient provides a measure of the validity of the model predictions relative to the actual data.

4. Results and discussions

Occurrence of geomagnetic storms in solar cycles

The table 1 provides a detailed analysis of the number of GmS in each solar cycle and in different phases of the cycles.

Generally, it is observed that the declining phases of solar cycles tend to have a higher frequency of geomagnetic storms compared to the ascending phases. The number of geomagnetic storms in the declining phases of solar cycles is nearly double that of the rising phases (see Table 1) is supported by several studies in the scientific literature (see for example (Abe et al., 2023) and Echer et al. (2013)) for a discusion about larger number of storms in the descending phase of solar cycle may be related to moderate storm occurrence). This can be attributed to several factors, such as the increased activity of coronal holes that emit high-speed solar wind streams, and the more favorable configuration of the interplanetary magnetic field for magnetic reconnection during the descending phases (Gonzalez et al., 1994).

During the declining phases of the solar cycle, long-lasting coronal holes are more common, which emit high-speed solar wind streams contributing to a higher frequency of geomagnetic storms. Additionally, in the descending phase, the orientation and structure of the interplanetary magnetic field tend to be more favorable for magnetic reconnection with the Earth's magnetic field, facilitating the conditions for geomagnetic storms (Verbanac et al., 2011). In other word, the IMF B_z component is a significant driver of geomagnetic activity due to the magnetic reconnection mechanism. When the IMF B_z is oriented southward, it interacts with the Earth's northward magnetic field at the dayside magnetopause, facilitating magnetic reconnection. This process allows solar wind energy to penetrate the magnetosphere, which can lead to enhanced geomagnetic activity and the development of geomagnetic storms (Gonzalez et al., 1994).

For instance, Richardson et al. (2001) note that the frequency of severe space weather events, including geomagnetic storms, tends to be higher during the descending phases of the solar cycle. Similar observations have been reported by Tsurutani et al. (1992, 1995) and also by Ji et al. (2012), who also found that the declining phases are associated with increased geomagnetic activity. These findings underscore the importance of considering the phase of the solar cycle when studying and predicting geomagnetic storm occurrences.

It is observed that, in general, the heliospheric variables show a more notable correlation with the number of geomagnetic storms compared to more common variables such as sunspot number or solar flux. A correlation coefficient for number of GmS in terms of predictor variables can be found in 2 providing an additional visualization of the correlation between the predictor variables and the total number of GmS, as well as their specific correspondence with the total number of storms in each class, from moderate to severe. Specifically, the correlation coefficients for PT, IMF Alpha/Proton ratio and sunspot are generally higher, indicating stronger (positive) relationships with the number of GmS. For instance, the correlation coefficients (CC) PT, IMF, A/P ratio and sunspot are 0.49, 0.71, 0.56 and 0.51, respectively. This observation suggests that solar wind and interplanetary magnetic field conditions may have a more direct influence on geomagnetic activity, however, sunspot has a high alignment with the number of solar storms assuming strong relationships. This finding highlights the importance of considering a wide range of heliophysical variables when studying and predicting geomagnetic activity, as these less conventional variables may provide a better understanding of the underlying mechanisms involved in GmS generation. It is observed that, as the intensity of the storms increases, the correlation with the predictor variables tends to systematically decrease. This finding suggests that predictor variables

Table 1. Statistics and count of total GmS for both phases of cycle for each cycle considered in this study.

Cycle	Phase	Start	End	Weak	Moderate	Strong	Severe	Extreme
20	rising	1964-10	1968-11	508	234	46	2	1
	declining	1968-11	1976-03	1585	603	85	1	0
21	rising	1976-03	1979-12	1074	496	61	0	0
	declining	1979-12	1986-10	2089	968	163	5	0
22	rising	1986-10	1989-11	964	501	68	4	2
	declining	1989-11	1996-08	2276	1228	203	6	0
23	rising	1996-08	2001-11	1236	525	121	4	0
	declining	2001-11	2008-12	1488	574	111	6	2
24	rising	2008-12	2014-04	604	232	21	0	0
	declining	2014-04	2019-12	880	262	25	6	1
25	rising	2019-12	2025-01	450	132	15	0	0

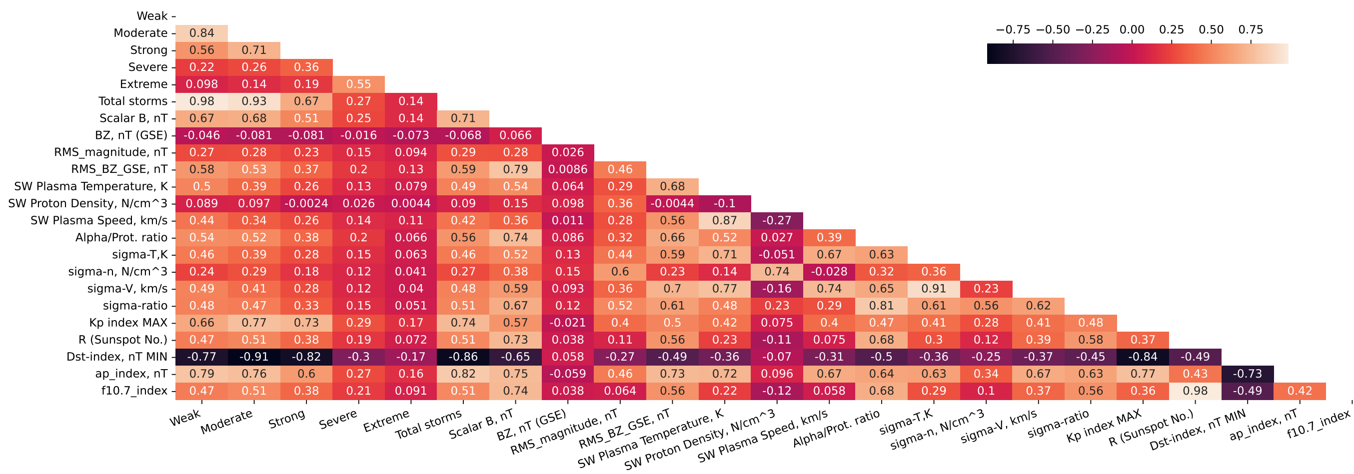

Fig. 2. Correlation coefficients between geomagnetic storms (GmS) and various predictor variables across all studied solar cycles and the complete dataset. The predictor variables include sunspot number, solar radio flux, proton temperature, proton density, plasma speed, alpha/proton ratio, flow pressure, interplanetary magnetic field strength, and the minimum Dst index. The correlation coefficients are depicted for different intensities of geomagnetic storms categorized as weak, moderate, strong, severe, and great. The standard deviations of each of the heliophysical variables have been included in the figure for comparison purposes.

may be more effective in predicting lower intensity GmS, while their predictive ability may be less accurate for higher magnitude storms. This pattern may have significant implications for the development of geomagnetic storm prediction models and highlights the importance of considering storm intensity when assessing the relationship with predictor variables.

However, for the total number of storms, which is represented as the sum of all storms independent of their intensity, the correlation is also considered, this has to do with the fact that the number of more intense storms is relatively much smaller than the smaller or moderate ones, so that the sum is dominated mainly by the storms of lower intensity.

Modeling geomagnetic storms from heliophysical predictors

The first regression model we employ is a linear multiple regression model, which allows us to explore linear relationships between the number of GmS and a set of predictor variables related to space weather. The predictor variables used in the model include solar activity represented by sunspot number and solar radio, as well as variables related to solar wind properties such as proton temperature and density, plasma velocity and pressure, alpha/proton ratio, and interplanetary magnetic field magnitude. Tables 2 shows the regression coefficients and their uncertainties. As can be seen from the table, except for the number of sunspots and Proton Density, all predictor variables are statistically significant with a confidence interval of $\alpha = 0.05$.

The regression coefficients provide information on the strength and direction of the relationship between each predictor variable and the number of GmS. For example, negative coefficients for temperature, Alpha/Proton ratio and Dst Min suggest an inverse relationship with the number of GmS, while positive coefficients for plasma speed and pressure, as well as magnetic field magnitude, indicate a direct relationship. The relationship between sunspots and the number of storms is more complex, with a negative coefficient for sunspots and a positive coefficient for solar radio activity, suggesting a nonlinear relationship between these variables and GmS.

The root mean squared error of this model, which indicates the discrepancy between observed and predicted values of GmS, is 63.041. This suggests that the model has moderate accuracy in predicting the number of GmS, although there is

Table 2. Parameters and uncertainties for multiple linear regression. R-squared: 0.671, Adj. R-squared: 0.668, Log-Likelihood: -4010.7, AIC: 8037, BIC: 8074.

Variables	**coef**	**std err**	t	$P > \|t\|$	**[0.025**	**0.975]**
const	-472.6417	46.042	-10.265	0.000*	-563.036	-382.248
Scalar B, nT	19.9998	3.751	5.333	0.000*	12.636	27.363
SW Plasma Temperature, K	-0.0003	0.000	-2.078	0.038*	-0.001	-1.92e-05
SW Proton Density, N/cm^3	8.9535	1.798	4.978	0.000*	5.423	12.484
SW Plasma Speed, km/s	0.8279	0.124	6.667	0.000*	0.584	1.072
Alpha/Prot. ratio	-533.3684	285.013	-1.871	0.062	-1092.933	26.196
R (Sunspot No.)	0.2786	0.063	4.395	0.000*	0.154	0.403
Dst-index, nT MIN	-0.7810	0.047	-16.509	0.000*	-0.874	-0.688

still room for improvement in model accuracy.

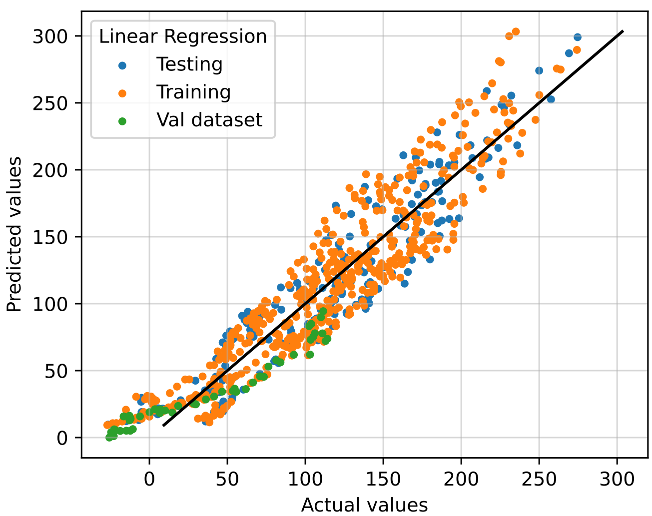

Fig. 3. Scatter plot between actual and predicted values for GmS in all solar cycles with multiple linear regression model. Coeficients in multiple linear model are displayed in Table 2.

We have generated scatter plots for the regression obtained between these parameters for the predictions and the actual values. The equation of the linear fit (in table 2) and the Pearson correlation coefficient is shown in Figure 3. The scatter plot between the GmS Number index and the predicted values is shown in Figure 3, indicating a high correlation coefficient of 0.815 but a determination coefficient of 0.668. It is important to note that the Pearson correlation coefficient is not always a reliable estimator of regression quality, as it only measures linear relationships and may not capture nonlinear associations that could be present in the data.

4.1. ML models

ML models play a key role in predicting and understanding complex geomagnetic phenomena. In this section, we analyze the results of applying several machine learning methods, including Random Forest regression, Gradient Boosting, Adaboost, and ExtraTree Regressor, to model and predict the number of GmS as a function of selected predictor variables. These models were trained using historical GmS data and heliophysical variables, with the goal of identifying meaningful patterns and relationships that can aid in the prediction of future GmS.

Before starting the training of the models, the data set covering solar cycles 20 to 24 (up to December 2019) was divided into two subsets: a training set (consisting of 70% of this data selected randomly) and a test set (comprising the remaining 30%). For validation purposes, we created an external validation set using data from cycle 25 (starting from December 2019), which is separate from both the whole set. This approach allows us to validate the model on data that the model has not encountered during training or testing, ensuring a more robust evaluation of its performance.

During the training and tuning process of the machine learning models, exhaustive hyperparameter searches were conducted using the `GridSearchCV` cross-validation strategy. For each model, the best estimators and hyperparameter combinations that minimized evaluation metrics such as RMSE were identified, resulting in optimal models for predicting the number of GmS. For example, the best pipeline for the Random Forest model included a `ccp_alpha` of 0.001, `criterion` set to Poisson, `max_depth` of 500, `max_features` as `log2`, `min_samples_leaf` of 1, `min_samples_split` of 2, and 1600 estimators. The Gradient Boosting model performed best with a `ccp_alpha` of 0.0001, `criterion` as Friedman MSE, a learning rate of 0.1, `loss` as squared error, `max_depth` of 500, `max_features` as `log2`, `min_samples_leaf` of 8, `min_samples_split` of 16, and 200 estimators. The `HistGradientBoostingRegressor` achieved optimal results with a learning rate of 0.1, `loss` set to Poisson, `max_bins` of 50, `max_depth` of 200, `max_features` as 0.7, and `min_samples_leaf` of 8. The best pipeline for the Extra Trees Regressor included `bootstrap` set to False, `ccp_alpha` of 0.0, `criterion` as squared error, `max_depth` of 50, `max_features` as 0.9, `min_samples_leaf` of 1, and `min_samples_split` of 2. Finally, the AdaBoost Regressor achieved optimal performance with a learning rate of 0.1, `loss` set to exponential, and 1800 estimators. These results highlight the importance of hyperparameter fitting in building accurate and efficient machine learning models for predicting geomagnetic activity.

The best model we have found is the Random Forest regressor model. This model has demonstrated excellent performance across several evaluation metrics (Table 3). For example, on the training set, the RF achieved an RMSE of 4.959 and a coefficient of determination (R2 score) of 0.994, suggesting that approximately 99.4% of the variability in the number of GmS

Table 3. Performance and evaluation of machine learning models on the training set and the full data set. The root mean square error (RMSE), coefficient of determination (r2-score), correlation coefficient, Rand-Score, Adj-Rand-Score, AIC and BIC are shown for each model. The best results in terms of RMSE, r2-score and correlation coefficient are highlighted in bold.

MODEL	SET	RMSE	R2 score	MAPE	CORRCOEF	RANDS	ADJ_RANDS	AIC	BIC
RF	Training	4.959	0.994	0.040	0.997	1.000	1.000	3058.271	3087.829
	TEST	11.563	0.964	0.087	0.983	1.000	1.000	1692.158	1715.817
	Full	7.578	0.985	0.054	0.993	1.000	1.000	4980.541	5012.606
	Validation	8.318	0.921	0.230	0.988	1.000	1.000	367.731	381.115
GB	Training	0.464	1.000	0.003	1.000	1.000	1.000	670.229	699.787
	Testing	10.578	0.970	0.076	0.986	1.000	1.000	1653.525	1677.184
	Full	5.816	0.992	0.025	0.996	1.000	1.000	4598.933	4630.997
	Validation	5.315	0.963	0.188	0.989	0.997	0.000	322.945	336.329
HGB	Training	2.344	0.999	0.019	0.999	1.000	0.000	2302.899	2332.457
	Testing	9.784	0.975	0.067	0.988	1.000	0.000	1619.654	1643.313
	Full	5.714	0.992	0.034	0.996	1.000	0.000	4573.399	4605.463
	Validation	5.675	0.960	0.177	0.985	0.981	0.000	329.509	342.893
ET	Training	0.000	1.000	0.000	1.000	1.000	1.000	-28195.723	-28166.165
	Testing	9.484	0.977	0.064	0.989	1.000	1.000	1606.133	1629.793
	Full	5.203	0.993	0.019	0.997	1.000	1.000	4438.256	4470.321
	Validation	4.933	0.967	0.173	0.990	1.000	1.000	315.490	328.874
AB	Training	13.144	0.954	0.141	0.980	0.999	0.000	4040.850	4070.408
	Testing	16.292	0.930	0.152	0.967	0.999	0.000	1840.965	1864.624
	Full	14.165	0.946	0.144	0.976	0.999	0.000	5882.530	5914.595
	Validation	18.716	0.528	0.401	0.966	0.944	0.000	448.833	462.217
LR	Training	62.403	0.509	0.575	0.816	0.993	0.000	5610.980	5640.538
	Testing	64.498	0.508	0.582	0.826	0.996	0.000	2438.140	2461.800
	Full	63.041	0.510	0.577	0.819	0.994	0.000	8035.454	8067.518
	Validation	39.337	0.607	1.345	0.799	0.986	0.000	523.111	536.496

can be explained by the model.

In addition, the Random Forest model exhibited a high correlation coefficient of 0.997 in the training set, indicating a strong linear relationship between the predictor variables and the target variable. This high level of correlation suggests that the model effectively captures the relationship between the input variables and the number of GmS.

The performance of the RF model was further validated on the full dataset, where it showed an RMSE of 7.578 and an R2 score of 0.985, indicating a robust ability to generalize to unseen data. Moreover, the correlation coefficient of 0.993 on the full dataset reinforces the predictive quality of the model. The RF model has proven to be the most effective for predicting the number of GmS based on the selected predictor variables, exhibiting a high level of accuracy and generalizability.

An important consideration in selecting the best model is the consistency of performance across different data sets: training, testing, full, and validation. The Random Forest model showed the smallest difference in metrics such as RMSE, R2 score, and correlation coefficient across these datasets. This suggests that the RF model is well-balanced, effectively capturing the underlying patterns in the data without overfitting or underfitting. The ability of the RF model to maintain stable performance across various subsets of data reinforces its suitability as the most reliable model for predicting the number of GmS, highlighting its robustness and adaptability.

Figure 4 shows the scatter plot between the actual and predicted values for each of the ML models tested, further exhibiting the high performance of GB over all the rest.

While evaluating the models, we also considered the MAPE to assess predictive accuracy in terms of percentage error. The Random Forest model demonstrated good performance with low MAPE values across the training (0.040), test (0.087), and full datasets (0.054). Although the MAPE increased in the validation set (0.230), the model's overall accuracy in the initial datasets underscores its effectiveness. This suggests that while there is some variability in percentage error with unseen data, the RF model remains a strong contender due to its consistent performance in most scenarios.

In the RF model, it is observed that the most important variable in the prediction of the number of GmS is the Dst-index MIN, with an importance of 30.1%. This is expected since the number of geomagnetic storms is derived from the Dst index. However, it is interesting to note that heliospheric variables such as the scalar B (26.5%) and the alpha/proton ratio (16.2%) occupy prominent positions in terms of importance (See Fig. 5). These three variables account for approximately 72.8% of the total importance in the model's prediction, indicating that they are the main drivers of the number of GmS according to the model. Additionally, the importance of the sunspot number R, with an importance of 10.3%, underscores the influence of solar activity on geomagnetic storms. Other significant variables include the solar wind plasma temperature (9.8%), solar wind proton density (4.5%), and solar wind plasma speed (2.5%), the first of them with almost as much importance as sunspot.

In contrast with findings from authors such as Abe et al. (2023), who identified a strong correlation between the number of GmS and sunspot activity, our study highlights the promi-

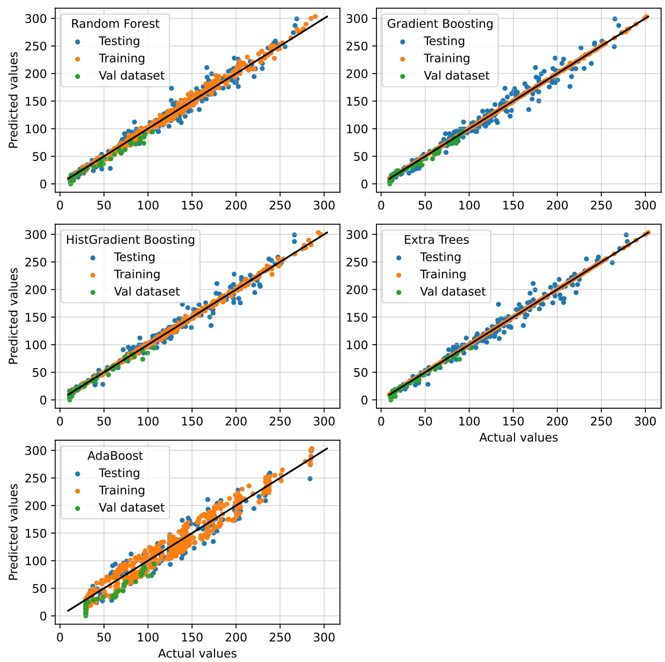

Fig. 4. Scatter plot of the actual and predicted values by the machine learning models. Each point on the graph represents a pair of values: the actual value of GmS on the y-axis, and the value predicted by the model on the x-axis. Points closer to the diagonal line represent a better prediction, as they indicate a smaller difference between the actual and predicted values.

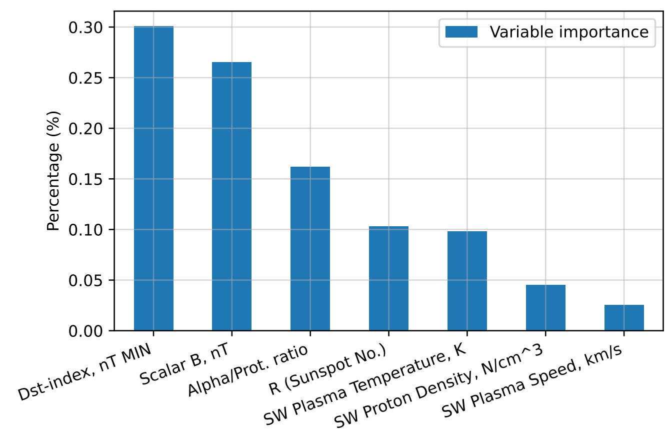

Fig. 5. Variable importance for the Random Forest model predicting the number of geomagnetic storms (GmS). The Dst-index MIN is the most influential variable, reflecting its direct relation to geomagnetic storm quantification. Heliophysical variables such as the scalar B, alpha/proton ratio, and sunspot number R also show significant importance, highlighting their role in influencing geomagnetic activity. Other variables, including solar wind plasma temperature, proton density, and plasma speed, contribute to the model's predictive capability.

nence of the Interplanetary Magnetic Field (IMF). This variance in results underscores the significance of alternative heliophysical variables such as the IMF, proton temperature, and flow pressure.

The importance of the mean IMF as a significant predictor of the number of GmS is supported by its interaction with the solar wind and the Earth's magnetic field. The IMF is a critical component of the solar wind that interacts with the Earth's magnetic field during geomagnetic events. Variations in the IMF can affect the Earth's magnetosphere, triggering GmS. Additionally, the IMF transports energy from the Sun to the Earth, and fluctuations in this field can influence the amount of energy transferred during space weather events. This energy transfer plays a crucial role in the generation and amplification of GmS (Kane, 2005; Gonzalez et al., 1999).

The IMF is a key indicator of solar wind conditions that can influence the Earth's magnetosphere. Variations in solar wind speed, density, and direction, all associated with the IMF, can trigger responses in the magnetosphere that lead to GmS. Therefore, the IMF not only acts as a predictor of geomagnetic storms but also highlights the dynamic interactions between solar and terrestrial environments.

On the other hand, variables such as proton temperature, alpha/proton ratio, proton density, sunspot number R, and F10.7 are of minor importance compared to the first three variables mentioned. However, they still contribute to the model, accounting for 10% of the total importance (Boroyev et al., 2020; Inyurt, 2020).

This analysis suggests that magnetic field-related features, plasma speed, and flow pressure are the most influential factors in predicting the number of GmS according to the Random Forest model. These findings may be useful to better understand the mechanisms behind GmS and to develop more accurate prediction strategies in the future.

The primary purpose of using different machine learning models in our study is to identify the most effective approach for predicting the number of geomagnetic storms based on heliospheric variables. By employing various models, including Random Forest regression, Gradient Boosting, AdaBoost, and ExtraTree Regressor, we can compare their performance using evaluation metrics such as RMSE, R2 score, and correlation coefficients. This comparison helps us identify the strengths and weaknesses of each model. Different models may capture different aspects of the data. For instance, some models may handle non-linear relationships better, while others may be more robust to outliers. By testing multiple models, we ensure that our final predictions are robust and reliable.

Each machine learning model has its own set of hyperparameters that can be fine-tuned to improve performance. By exploring various models, we can determine the optimal hyperparameter settings for each, leading to better predictive accuracy. Additionally, different models provide different methods for assessing the importance of predictor variables. By using multiple models, we can cross-validate the importance of key variables such as the Interplanetary Magnetic Field (IMF), plasma speed, and proton temperature, gaining deeper insights into their roles in geomagnetic storm prediction.

The application of various machine learning models in our study highlights the importance of robust model comparison and hyperparameter optimization in developing accurate predictive models for geomagnetic storms. The findings emphasize the significance of heliospheric variables in influencing geomagnetic activity and underscore the need for a comprehensive approach in studying and predicting space weather events. The insights gained from this research can aid in the development of more effective prediction strategies, contributing to better preparedness and mitigation of the impacts of geomagnetic storms.

5. Conclusions

After an analysis of the data and the application of various machine learning models to predict the number of GmS, several significant conclusions can be drawn:

- The study shows that the number of geomagnetic storms (GmS) tends to be slightly higher in odd-numbered solar cycles compared to even-numbered cycles. This pattern aligns with the 22-year cycle of geomagnetic activity described by Cliver et al. (1996). According to this cycle, peaks in geomagnetic activity alternate in strength between odd and even solar cycles due to the reversal of the solar magnetic field's polarity. This phenomenon results in more intense geomagnetic activity during certain phases of the 22-year cycle, contributing to the observed differences in storm counts between odd and even solar cycles.

- In addition to the cycle-based variations, our data reveal that the number of GmS is significantly higher during the downward phases of solar cycles compared to the upward phases. This trend is evident across all analyzed solar cycles and is supported by our visual and tabular analysis. The increased geomagnetic activity during the declining phases can be attributed to the presence of high-speed solar wind streams and favorable IMF Bz conditions, which facilitate magnetic reconnection processes that drive geomagnetic storms. This aligns with the findings of Gonzalez et al. (1994), who emphasized the critical role of southward IMF Bz in geomagnetic activity.

- Certain variables, such as mean interplanetary magnetic field, plasma speed and flow pressure, have been found to have a significant correlation with the number of GmS. These variables emerged as the main drivers in predicting the number of storms according to the machine learning models used.

- Random Forest models proved to be the most effective in predicting the number of GmS, with lower RMSE and higher coefficient of determination (R2 score) compared to other models. This suggests that Random Forest is the most robust and accurate approach for this type of prediction in our dataset.

- Analysis of the importance of predictor variables revealed that the Dst-index MIN, scalar B, and alpha/proton ratio are the most influential factors in predicting the number of GmS. These findings provide valuable information on the underlying mechanisms that drive GmS and may guide future research in this field.

This study we intent to contribute to a better understanding of the behavior of GmS and has demonstrated the effectiveness of machine learning models, in predicting this phenomenon. These findings have important implications for the prediction and mitigation of the adverse effects of GmS on Earth and technological infrastructures sensitive to space weather variations.

Acknowledgments

The authors would like to thank the UTB Research Department for its unconditional support in this research. We would like to extend our sincere gratitude to the two anonymous referees for their invaluable suggestions and thorough discussions. Their insightful feedback has significantly contributed to the revision and enhancement of our conclusions.

References

Abe, O., Fakomiti, M., Igboama, W. et al. (2023). Statistical analysis of the occurrence rate of geomagnetic storms during solar cycles 20–24. *Advances in Space Research*, *71*(5), 2240–2251. doi:https://doi.org/10.1016/j.asr.2022.10.033.

Abraha, G., Yemane, T., & Kassa, T. (2020). Geomagnetic storms and their impacts on ethiopian power grid. *Advances in Astronomy and Space Physics*, *10*, 55–64.

Adnan, M., Alarood, A. A. S., Uddin, M. I. et al. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Computer Science*, *8*, e803. doi:https://doi.org/10.7717/peerj-cs.803.

Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. In *2021 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1551–1559). IEEE. doi:https://doi.org/10.1109/CEC45853.2021.9504761.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967. doi:https://doi.org/10.1007/s10462-020-09896-5.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197–227. doi:https://doi.org/10.1007/s11749-016-0481-7.

Borovsky, J. E., & Shprits, Y. Y. (2017). Is the dst index sufficient to define all geospace storms? *Journal of Geophysical Research: Space Physics*, *122*(11), 11–543. doi:https://doi.org/10.1002/2017JA024679.

Boroyev, R. N., Vasiliev, M. S., & Baishev, D. G. (2020). The relationship between geomagnetic indices and the interplanetary medium parameters in magnetic storm main phases during cir and icme events. *Journal of Atmospheric and Solar-Terrestrial Physics*, *204*, 105290. doi:https://doi.org/10.1016/j.jastp.2020.105290.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. doi:https://doi.org/10.1023/A:1010933404324.

Chiu, M., Von-Mehlem, U., Willey, C. et al. (1998). Ace spacecraft. *Space science reviews*, *86*, 257–284. doi:https://doi.org/10.1023/A:1005002013459.

Cliver, E. W., Boriakoff, V., & Bounar, K. H. (1996). The 22-year cycle of geomagnetic and solar wind activity. *Journal of Geophysical Research: Space Physics*, *101*(A12), 27091–27109. doi:https://doi.org/10.1029/96JA02037.

Drucker, H. (1997). Improving regressors using boosting techniques. In *Icml* (pp. 107–115). Citeseer volume 97. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6d8226a52ebc70c8d97ccae10a74e1b0a3908ec1.

Echer, E., Gonzalez, W. D., Tsurutani, B. T. et al. (2008). Interplanetary conditions causing intense geomagnetic storms (dst≤- 100 nt) during solar cycle 23 (1996–2006). *Journal of Geophysical Research: Space Physics*, *113*(A5). doi:https://doi.org/10.1029/2007JA012744.

Echer, E., Tsurutani, B., & Gonzalez, W. (2013). Interplanetary origins of moderate (- 100 nt< dst≤- 50 nt) geomagnetic storms during solar cycle 23 (1996–2008). *Journal of Geophysical Research: Space Physics*, *118*(1), 385–392. doi:https://doi.org/10.1029/2012JA018086.

Eid, A., Nawawy, M., & Robaa, S. (2022). Geomagnetic storm impacts on communication, navigation, surveillance, and air traffic management (cns/atm). *Current Science International*, *11*(3), 282–290.

Fairfield, D., Lepping, R., Hones Jr, E. et al. (1981). Simultaneous measurements of magnetotail dynamics by imp spacecraft. *Journal of Geophysical Research: Space Physics*, *86*(A3), 1396–1414. doi:https://doi.org/10.1029/JA086iA03p01396.

Frank, L. A. (1994). *Comprehensive Plasma Instrumentation (CPU) for the Geotail spacecraft*. Technical Report. URL: https://ntrs.nasa.gov/api/citations/19960041496/downloads/19960041496.pdf.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378. doi:https://doi.org/10.1016/S0167-9473(01)00065-2.

Garrard, T., Davis, A., Hammond, J. et al. (1998). The ace science center. *The Advanced Composition Explorer Mission*, (pp. 649–663). doi:https://doi.org/10.1007/978-94-011-4762-0_23.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42. doi:https://doi.org/10.1007/s10994-006-6226-1.

Gonzalez, W., Joselyn, J.-A., Kamide, Y. et al. (1994). What is a geomagnetic storm? *Journal of Geophysical Research: Space Physics*, *99*(A4), 5771–5792. doi:https://doi.org/10.1029/93JA02867.

Gonzalez, W. D., Echer, E., Tsurutani, B. T. et al. (2011). Interplanetary origin of intense, superintense and extreme geomagnetic storms. *Space science reviews*, *158*, 69–89. doi:https://doi.org/10.1007/s11214-010-9715-2.

Gonzalez, W. D., Tsurutani, B. T., & Clúa de Gonzalez, A. L. (1999). Interplanetary origin of geomagnetic storms. *Space Science Reviews*, *88*(3-4), 529–562. doi:https://doi.org/10.1023/A:1005160129098.

Hady, A. (2009). Descriptive study of solar activity sudden increase and halloween storms of 2003. *Journal of atmospheric and solar-terrestrial physics*, *71*(17-18), 1711–1716. doi:https://doi.org/10.1016/j.jastp.2008.11.019.

Hajra, R., Marques de Souza Franco, A., Echer, E. et al. (2021). Long-term variations of the geomagnetic activity: A comparison between the strong and weak solar activity cycles and implications for the space climate. *Journal of Geophysical Research: Space Physics*, *126*(4), e2020JA028695. doi:https://doi.org/10.1029/2020JA028695.

Hall, C. M., & Johnsen, M. G. (2020). Possible influence of variations in the geomagnetic field on migration paths of snow buntings. *International Journal of Astrobiology*, *19*(2), 195–201. doi:https://doi.org/10.1017/S1473550419000193.

Hastie, T., Tibshirani, R., Friedman, J. et al. (2009). Random forests. *The elements of statistical learning: Data mining, inference, and prediction*, (pp. 587–604). doi:https://doi.org/10.1007/978-0-387-84858-7_15.

Inyurt, S. (2020). Modeling and comparison of two geomagnetic storms. *Advances in Space Research*, *65*(3), 966–977. doi:https://doi.org/10.1016/j.asr.2019.11.004.

Ji, E.-Y., Moon, Y.-J., Gopalswamy, N. et al. (2012). Comparison of dst forecast models for intense geomagnetic storms. *Journal of Geophysical Research: Space Physics*, *117*(A3). doi:https://doi.org/10.1029/2011JA016872.

Kane, R. P. (2005). How good is the relationship of solar and interplanetary plasma parameters with geomagnetic storms? *Journal of Geophysical Research: Space Physics*, *110*(A2). doi:https://doi.org/10.1029/2004JA010799.

Kiznys, D., Vencloviene, J., & Milvidaitė, I. (2020). The associations of geomagnetic storms, fast solar wind, and stream interaction regions with cardiovascular characteristic in patients with acute coronary syndrome. *Life Sciences in Space Research*, *25*, 1–8. doi:https://doi.org/10.1016/j.lssr.2020.01.002.

Lakhina, G. S., & Tsurutani, B. T. (2016). Geomagnetic storms: historical perspective to modern view. *Geoscience Letters*, *3*, 1–11. doi:https://doi.org/10.1186/s40562-016-0037-4.

Le, G.-M., Cai, Z.-Y., Wang, H.-N. et al. (2013). Solar cycle distribution of major geomagnetic storms. *Research in Astronomy and Astrophysics*, *13*(6), 739. doi:https://doi.org/10.1088/1674-4527/13/6/013.

Loewe, C., & Prölss, G. (1997). Classification and mean behavior of magnetic storms. *Journal of Geophysical Research: Space Physics*, *102*(A7), 14209–14213. doi:https://doi.org/10.1029/96JA04020.

Lopez, R. E., Baker, D. N., & Allen, J. (2004). Sun unleashes halloween storm. *Eos, Transactions American Geophysical Union*, *85*(11), 105–108. doi:https://doi.org/10.1029/2004EO110002.

Love, J. J., Rigler, E. J., Hartinger, M. D. et al. (2023). The march 1940 superstorm: Geoelectromagnetic hazards and impacts on american communication and power systems. *Space Weather*, *21*(6), e2022SW003379. doi:https://doi.org/10.1029/2022SW003379.

Mandea, M., & Chambodut, A. (2020). Geomagnetic field processes and their implications for space weather. *Surveys in Geophysics*, *41*, 1611–1627. doi:https://doi.org/10.1007/s10712-020-09598-1.

Miteva, R., Nedal, M., Samwel, S. W. et al. (2023). Parameter study of geomagnetic storms and associated phenomena: Cme speed de-projection vs. in situ data. *Universe*, *9*(4), 179. doi:https://doi.org/10.3390/universe9040179.

Moore, J. H., Ribeiro, P. H., Matsumoto, N. et al. (2023). Genetic programming as an innovation engine for automated machine learning: The tree-based pipeline optimization tool (tpot). In *Handbook of Evolutionary Machine Learning* (pp. 439–455). Springer. doi:https://doi.org/10.1007/978-981-99-3814-8_14.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21. doi:https://doi.org/10.3389/fnbot.2013.00021.

Nitta, N. V., Mulligan, T., Kilpua, E. K. et al. (2021). Understanding the origins of problem geomagnetic storms associated with "stealth" coronal mass ejections. *Space Science Reviews*, *217*(8), 82. doi:https://doi.org/10.1007/s11214-021-00857-0.

Nti, I. K., Nyarko-Boateng, O., Aning, J. et al. (2021). Performance of machine learning algorithms with different k values in k-fold cross-validation. *International Journal of Information Technology and Computer Science*, *13*(6), 61–71. doi:https://doi.org/10.5815/ijitcs.2021.06.05.

Ogilvie, K., & Desch, M. (1997). The wind spacecraft and its early scientific results. *Advances in Space Research*, *20*(4-5), 559–568. doi:https://doi.org/10.1016/S0273-1177(97)00439-0.

Olson, R. S., & Moore, J. H. (2016). Tpot: A tree-based pipeline optimization tool for automating machine learning. In *Workshop on automatic machine learning* (pp. 66–74). PMLR. doi:https://doi.org/10.1007/978-3-030-05318-5_8.

Paularena, K., & King, J. (1999). Nasa's imp 8 spacecraft. *Interball in the ISTP Program: Studies of the solar wind-magnetosphere-ionosphere interaction*, (pp. 145–154). doi:https://doi.org/10.1007/978-94-011-4487-2_11.

Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/.

Rathore, B., Kaushik, S., Bhadoria, R. et al. (2012). Sunspots and geomagnetic storms during solar cycle-23. *Indian Journal of Physics*, *86*, 563–567. doi:https://doi.org/10.1007/s12648-012-0106-2.

Reyes, P. I., Pinto, V. A., & Moya, P. S. (2021). Geomagnetic storm occurrence and their relation with solar cycle phases. *Space Weather*, *19*(9), e2021SW002766. doi:https://doi.org/10.1029/2021SW002766.

Richardson, I., Cliver, E., & Cane, H. (2001). Sources of geomagnetic storms for solar minimum and maximum conditions during 1972–2000. *Geophysical Research Letters*, *28*(13), 2569–2572. doi:https://doi.org/10.1029/2001GL013052.

Samwel, S., & Miteva, R. (2023). Correlations between space weather parameters during intense geomagnetic storms: Analytical study. *Advances in Space Research*, *72*(8), 3440–3453. doi:https://doi.org/10.1016/j.asr.2023.07.053.

Sarimov, R. M., Serov, D. A., & Gudkov, S. V. (2023). Biological effects of magnetic storms and elf magnetic fields. *Biology*, *12*(12), 1506. doi:https://doi.org/10.3390/biology12121506.

Schaffer, C. (1993). Selecting a classification method by cross-validation. *Machine learning*, *13*, 135–143. doi:https://doi.org/10.1007/BF00993106.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*, (pp. 149–171). doi:https://doi.org/10.1007/978-0-387-21579-2_9.

Schmidt, R., Arends, H., Pedersen, A. et al. (1995). Results from active spacecraft potential control on the geotail spacecraft. *Journal of Geophysical Research: Space Physics*, *100*(A9), 17253–17259. doi:https://doi.org/10.1029/95JA01552.

Singh Chauhan, D., Tiwari, D., Tripathi, A. et al. (2010). Association of large geomagnetic storms with halo cmes and cirs observed during 1997–2007. *Indian Journal of Physics*, *84*, 881–886. doi:https://doi.org/10.1007/s12648-010-0052-9.

Siscoe, G., Crooker, N., & Clauer, C. (2006). Dst of the carrington storm of 1859. *Advances in Space Research*, *38*(2), 173–179. doi:https://doi.org/10.1016/j.asr.2005.02.102.

Srivastava, N., & Venkatakrishnan, P. (2002). Relationship between cme speed and geomagnetic storm intensity. *Geophysical research letters*, *29*(9), 1–1. doi:https://doi.org/10.1029/2001GL013597.

Sugiura, M. (1960). The average morphology of geomagnetic stroms with sudden commencements. *Abh. Akad. Wiss. Gottingen, Math.-phys.*, *53*.

Taran, S., Alipour, N., Rokni, K. et al. (2023). Effect of geomagnetic storms on a power network at mid latitudes. *Advances in Space Research*, *71*(12), 5453–5465. doi:https://doi.org/10.1016/j.asr.2023.02.027.

Tsurutani, B. T., Gonzalez, W. D., Gonzalez, A. L. et al. (1995). Interplanetary origin of geomagnetic activity in the declining phase of the solar cycle. *Journal of Geophysical Research: Space Physics*, *100*(A11), 21717–21733. doi:https://doi.org/10.1029/95JA01476.

Tsurutani, B. T., Gonzalez, W. D., Lakhina, G. et al. (2003). The extreme magnetic storm of 1–2 september 1859. *Journal of Geophysical Research: Space Physics*, *108*(A7). doi:https://doi.org/10.1029/2002JA009504.

Tsurutani, B. T., Gonzalez, W. D., Tang, F. et al. (1992). Great magnetic storms. *Geophysical Research Letters*, *19*(1), 73–76. doi:https://doi.org/10.1029/91GL02783.

Verbanac, G., Vršnak, B., Veronig, A. et al. (2011). Equatorial coronal holes, solar wind high-speed streams, and their geoeffectiveness. *Astronomy & astrophysics*, *526*, A20. doi:https://doi.org/10.1051/0004-6361/201014617.

Wilson III, L. B., Brosius, A. L., Gopalswamy, N. et al. (2021). A quarter century of wind spacecraft discoveries. *Reviews of Geophysics*, *59*(2). doi:https://doi.org/10.1029/2020RG000714.

Yacouba, S., Somaïla, K., & Louis, Z. J. (2022). Factors of geomagnetic storms during the solar cycles 23 and 24: A comparative statistical study. *Scientific Research and Essays*, *17*(3), 46–56. doi:https://doi.org/10.5897/SRE2022.6751.

Ying, C., Qi-Guang, M., Jia-Chen, L. et al. (2013). Advance and prospects of adaboost algorithm. *Acta Automatica Sinica*, *39*(6), 745–758. doi:https://doi.org/10.1016/S1874-1029(13)60052-X.

Zhang, S., He, L., & Wu, L. (2020). Statistical study of loss of gps signals caused by severe and great geomagnetic storms. *Journal of Geophysical Research: Space Physics*, *125*(9), e2019JA027749. doi:https://doi.org/10.1029/2019JA027749.