International Journal of
**Physical and Chemical Sciences**
https://www.ioro.org/ijpcs

IORO

# Prediction of Geomagnetic Storms through Artificial Intelligence: A Review, Methodological Framework, and Simulated Case Study

**Divyansh Mishra[1], Rajesh Kumar Mishra[2] and Rekha Agarwal[3]**

[1]*XLRI – Xavier School of Management, Jamshedpur, Jharkhand 831001, India*

[2]*ICFRE–Tropical Forest Research Institute (Ministry of Environment, Forests & Climate Change, Govt. of India), P.O. RFRC, Mandla Road, Jabalpur, MP-482021, India*

[3]*Government Science College, Jabalpur, MP, India–482 001*

E-mail: divyanshspps@gmail.com, rajeshkmishra20@gmail.com, rajesh.mishra0701@gov.in,rekhasciencecollege@gmail.com

**Abstract**

*Geomagnetic storms, driven primarily by coronal mass ejections (CMEs) and high-speed solar wind streams interacting with the Earth's magnetosphere, pose significant risks to satellite operations, power grid infrastructure, aviation, and communication systems. Traditional empirical and physics-based models of geomagnetic indices such as Dst and Kp, while physically interpretable, often struggle to capture the highly nonlinear and non-stationary dynamics of solar wind-magnetosphere coupling, particularly during the storm recovery phase. Artificial Intelligence (AI) techniques, including artificial neural networks, ensemble tree-based methods, and multilayer perceptron (MLP) models, have demonstrated considerable promise for improving the accuracy and lead time of geomagnetic storm forecasts. This paper (i) reviews the evolution of AI-based approaches to Dst/Kp prediction; (ii) proposes a consolidated methodological framework for building and validating AI-driven space-weather forecasting systems; and (iii) reports a complete, reproducible experimental case study in which a physically-grounded Burton-McPherron-Russell (BMR) ring-current simulator was used to generate a multi-year hourly solar-wind-driven Dst dataset, on which Random Forest, Gradient Boosting, Multilayer Perceptron, and linear baseline models were trained and evaluated under a chronologically leakage-aware train/validation/test protocol at 1-, 3-, and 6-hour forecast horizons. The best-performing model (MLP) achieved RMSE = 3.82 nT and $R^2$ = 0.956 at the 1-hour horizon, degrading to RMSE = 10.27 nT and $R^2$ = 0.684 at 6 hours, consistent with the general skill-decay pattern reported in the literature for physically driven Dst forecasting. Feature importance analysis confirmed that the most recent lagged Dst value and the 6-hour rolling mean of the southward IMF component (Bz) dominate predictive skill. Key challenges relating to data leakage, class imbalance for extreme events, multimodal data fusion, and model interpretability are discussed, along with directions for future research including physics-informed machine learning and probabilistic forecasting.*

**Keywords:** *geomagnetic storms; artificial intelligence; machine learning; Dst index; Burton-McPherron-Russell model; random forest; neural networks; space weather*

## 1. Introduction

Geomagnetic storms are transient disturbances of the Earth's magnetosphere caused by enhanced energy and momentum transfer from the solar wind, most commonly triggered by Earth-directed coronal mass ejections (CMEs) and corotating interaction regions (CIRs) associated with high-speed solar wind streams [1]. When the interplanetary magnetic field (IMF) turns strongly southward, magnetic reconnection at the dayside magnetopause intensifies, injecting energy into the ring current and producing measurable perturbations in the horizontal component of the geomagnetic field. These events can disrupt satellite electronics, degrade GNSS positioning accuracy, induce geomagnetically induced currents (GICs) in power transmission networks, and expose high-altitude aviation and astronauts to elevated radiation levels.

The severity of a geomagnetic storm is conventionally quantified using indices such as the Disturbance Storm Time (Dst) index, derived from four low-latitude magnetometer stations [2], and the planetary Kp index, derived from a global network of mid-latitude observatories. Because these indices respond nonlinearly to solar wind driving conditions and exhibit strong history-dependence, accurate short-term and medium-range forecasting has remained a long-standing challenge in space weather research.

Classical empirical models, beginning with the Burton-McPherron-Russell (BMR) relation linking solar wind parameters to Dst [3], and later refinements incorporating ring current decay and injection terms, provided the first quantitative forecasting tools. However, these physics-based formulations rely on simplifying assumptions that limit their accuracy during intense and superstorm events, and particularly during the storm recovery phase, which has repeatedly been shown to be poorly reproduced by linear and quasi-linear models [4]. Over the past three decades, Artificial Intelligence (AI) and machine learning (ML) methods have emerged as a complementary, data-driven paradigm capable of learning the complex, nonlinear input-output mapping between upstream solar wind measurements and geomagnetic response, without requiring an explicit physical parameterization.

This paper has three objectives. First, it reviews the trajectory of AI-based geomagnetic storm prediction research, from early feed-forward neural network models to contemporary deep learning architectures. Second, it proposes a consolidated methodological framework, covering data sources, feature engineering, model selection, and validation protocols, intended to guide the design of robust AI-based space weather forecasting systems. Third, and distinctively, it reports a complete, reproducible experimental pipeline: because live access to the NASA OMNIWeb / Kyoto World Data Center archives was not available within the computational environment used for this study, a physically-grounded synthetic dataset was generated using the BMR ring-current formalism driven by stochastic solar-wind forcing, and real machine learning models were trained, validated, and tested on this dataset end-to-end. This approach is explicitly disclosed as a simulation-based case study rather than an analysis of the observational archive, and is intended to demonstrate the proposed methodology on data with realistic storm statistics and known ground-truth dynamics, a strategy also used in the machine-learning literature for controlled algorithm benchmarking [12].

### 1.1. *Societal and Economic Impact*

The practical motivation for accurate, timely geomagnetic storm forecasting is well documented by historical impact events. The intense storm of 13–14 March 1989, associated with a Dst minimum near –589 nT, induced geomagnetically induced currents in the Hydro-Québec power transmission network sufficient to trip protective relays and collapse the entire grid within approximately 90 seconds, leaving roughly six million customers without electricity for up to nine hours. More recently, on 3 February 2022, SpaceX launched 49 Starlink satellites into a low deployment orbit; a moderate (NOAA G2-class) geomagnetic storm the following day enhanced thermospheric density and drag sufficiently that 38 of the 49 satellites re-entered the atmosphere within days, at an estimated cost of tens of millions of dollars [19]. Notably, this event involved only a moderate storm, illustrating that even non-extreme geomagnetic disturbances can have significant technological and economic consequences when they coincide with vulnerable operational windows, and reinforcing the operational value of short-horizon (1–6 hour) forecasts of the kind evaluated in Section 3 of this paper.

Beyond these headline events, sustained low-level geomagnetic activity contributes cumulative degradation to high-frequency radio communication, GNSS-based precision agriculture and surveying, and pipeline corrosion monitoring, while extreme (Carrington-class) events, last observed in September 1859 and considered plausible on decadal-to-centennial timescales, are estimated by some risk assessments to carry potential economic impact in the hundreds of billions to low trillions of US dollars if they were to occur in the present technologically dependent era. The so-called Halloween storms of October–November 2003, among the most intense of the space age (Dst minimum near −383 nT), caused a regional power outage in southern Sweden, forced airlines to reroute polar flights to avoid radiation exposure and HF communication blackouts, and damaged or degraded several Earth-observation and communications satellites, illustrating the multi-sector nature of geomagnetic storm risk.

This risk profile underlies the operational mandate of agencies including NOAA's Space Weather Prediction Center (United States), the UK Met Office Space Weather Operations Centre, the European Space Agency's Space Weather Service Network, and India's national space weather monitoring efforts, all of which increasingly incorporate, or are actively researching, machine-learning post-processing and forecasting components alongside their traditional physics-based and empirical models. This growing operational interest underlies the review of AI-based approaches conducted in Section 2.2, and motivates AI-based approaches capable of extracting maximal predictive skill from available solar wind monitoring assets.

## 2. Material and Methods

### 2.1. *Geomagnetic Indices and the Forecasting Problem*

The Dst index measures the average depression of the horizontal magnetic field at the equator caused by the ring current, and is reported at hourly resolution by the World Data Center for Geomagnetism, Kyoto [2]. Storms are conventionally classified by minimum Dst value as weak (−30 to −50 nT), moderate (−50 to −100 nT), intense (−100 to −200 nT), and superstorms (below −200 nT). The Kp index, an aggregate 3-hour planetary measure derived from thirteen mid-latitude observatories, is widely used operationally because of its global availability and its role in space weather alert thresholds issued by agencies such as NOAA's Space Weather Prediction Center.

From a forecasting standpoint, the problem is typically posed as a supervised time-series regression or sequence-to-sequence task: given a window of upstream solar wind measurements (from spacecraft such as ACE, DSCOVR, or Wind, positioned near the L1 Lagrange point) and/or recent index history, predict the index value one to several hours, or several days, ahead. The chaotic, non-stationary nature of the driving solar wind, combined with the roughly 30–60 minute propagation lag from L1 to the magnetosphere, makes this an inherently difficult nonlinear time-series problem, well suited to AI methods capable of capturing long-range temporal dependencies and nonlinear interactions among input variables.

The empirical basis for most AI feature sets remains the Burton-McPherron-Russell injection/decay picture [3], in which a ring-current index Dst* evolves according to $d(Dst^*)/dt = Q(t) - Dst^*/\tau$, where $Q(t)$ is an energy injection function driven by the southward component of the interplanetary electric field and $\tau$ is a ring-current decay time constant (typically several hours). The observed Dst is then related to Dst* through a dynamic-pressure correction term, $Dst = Dst^* + b\sqrt{P} - c$. This physically motivated structure underlies both classical empirical forecasting and the synthetic dataset generator described in Section 2.3 of this study.

Physically, a geomagnetic storm progresses through three canonical phases. The initial phase, lasting from minutes to a few hours, is associated with a sudden storm commencement (SSC): a sharp, positive jump in Dst caused by the sudden compression of the magnetopause when an interplanetary shock (often the leading edge of a CME-driven sheath) impinges on the magnetosphere, momentarily enhancing the Chapman-Ferraro magnetopause current system. The main phase follows once the IMF Bz turns and remains southward for a sustained interval, enabling continuous dayside reconnection that injects energetic particles into the ring current, deepening Dst over a period of typically 2–12 hours to its minimum value. The recovery phase, which can last from several hours to several days, reflects the gradual decay of the ring current through charge-exchange losses, wave-particle interactions, and other loss processes; it is this phase, with its multiple, physically distinct decay timescales, that classical single-exponential models such as the original BMR formulation reproduce least accurately, and which motivated much of the early neural-network literature reviewed in Section 2.2 [7], [8].

In addition to Dst, the SYM-H index provides a higher time-resolution (1-minute) analogue derived from a similar but expanded magnetometer network, and is increasingly used as an AI model target where sub-hourly forecast granularity is required. The auroral electrojet indices (AE, AU, AL), derived from high-latitude magnetometers, capture the substorm-related component of geomagnetic activity, which is dynamically coupled to, but not identical with, ring-current-driven Dst variations; some AI architectures reviewed in Section 2.2 incorporate AE as an auxiliary predictive feature or joint training target to better capture the substorm-storm coupling relevant to rapid Dst fluctuations during the main phase.

Operationally, geomagnetic storm forecasting depends on continuous monitoring of the solar wind upstream of Earth's bow shock. The Advanced Composition Explorer (ACE, launched 1997) and Deep Space Climate Observatory (DSCOVR, launched 2015) spacecraft, stationed near the Sun-Earth L1 Lagrange point approximately 1.5 million km sunward of Earth, provide the primary real-time plasma and magnetic field measurements used both operationally by agencies such as NOAA's Space Weather Prediction Center and as AI model inputs in the literature reviewed below. Because L1 lies upstream of Earth, these measurements provide a natural, if short (typically 30–60 minutes, depending on solar wind speed), forecast lead time before the corresponding disturbance reaches the magnetosphere — a lead time that AI models aim to extend through learned temporal structure rather than pure propagation delay alone.

***Table 1. Principal spacecraft and archives used as AI model inputs in the reviewed literature.***

| Source | Role | Location / coverage | Typical resolution |
|---|---|---|---|
| ACE | Solar wind plasma and IMF | Sun-Earth L1 (≈235 $R_e$ sunward) | 1–64 s (archived at 1 min–1 h) |
| DSCOVR | Solar wind plasma and IMF (operational backup/primary to ACE) | Sun-Earth L1 | 1 min–1 h |
| Wind | Solar wind plasma and IMF, multi-point with ACE/DSCOVR | L1 / near-Earth orbits | 3 s–1 h |
| OMNI (NASA/GSFC) | Merged, time-shifted, quality-controlled multi-spacecraft solar wind + indices | Propagated to bow shock nose | 1 min, 1 h, 1 day, 27 day |
| Kyoto WDC Dst/SYM-H | Ground-based low-latitude magnetometer indices | 4 low-latitude stations | 1 h (Dst), 1 min (SYM-H) |
| SDO/AIA | Extreme-ultraviolet solar imagery (precursor/multimodal input) | Geosynchronous / L1-independent | 10–12 s per wavelength channel |

## 2.2. *Review of AI-Based Approaches: Prior Literature*

The earliest applications of neural networks to geomagnetic storm prediction date to the mid-1990s. Wu and Lundstedt [6] applied Elman recurrent neural networks to predict geomagnetic storms from solar wind data, demonstrating that recurrent architectures with feedback connections could better capture the temporal evolution of storm development than static feed-forward networks. Gleisner, Lundstedt, and Wintoft [7] extended this line of work using time-delay neural networks to predict storm indices directly from upstream solar wind measurements.

Kugblenu, Taguchi, and Okuzawa [8] subsequently developed a feed-forward artificial neural network specifically targeting the Dst recovery phase, which earlier statistical models had consistently reproduced poorly. Their model achieved a correlation coefficient exceeding 0.95 between predicted and observed Dst during the recovery phase, with roughly 90% of observed Dst variance explained, and indicated that the minimum Dst reached during a storm is itself a significant predictor of the subsequent recovery trajectory. Lethy et al. [9] later extended neural-network Dst prediction with a systematic analysis of dependence on individual solar wind parameters.

As larger, cleaner solar wind datasets became available through missions such as ACE and OMNI, ensemble learning methods gained traction. Alobaid et al. [10] applied Random Forest and Gradient Boosting regressors to multi-solar-cycle Dst and heliospheric datasets spanning several decades, finding that these tree-based ensembles outperform traditional empirical formulations, and that the minimum Dst value, the interplanetary magnetic field magnitude, and the alpha-to-proton density ratio rank among the most influential predictive features, challenging earlier assumptions that sunspot activity alone was the dominant driver.

Support Vector Machines combined with distance-correlation-based feature selection have similarly been benchmarked against neural network counterparts for intense storm prediction. Lazzus et al. [11] applied Gaussian Process Regression, a Bayesian non-parametric approach, to intense geomagnetic storm forecasting using several decades of storm events, reporting root-mean-square errors around 19 nT and correlation coefficients near 0.87 for Dst prediction, while additionally providing calibrated predictive uncertainty, a feature that purely deterministic models lack.

The last decade has seen a shift toward deep learning architectures capable of modeling longer-range temporal dependencies. Yuan et al. [12] used a bagging ensemble of neural predictors for Dst prediction; Bergin, Chapman, and Watkins [13] examined deep feed-forward networks with carefully engineered training/validation splits; and Zhelavskaya et al. [14] reported multi-fidelity boosted neural networks for multi-hour-ahead Dst forecasting, all improving on classical baselines, particularly for the one-to-six-hour forecast horizon that is most operationally relevant for satellite operators and grid managers.

A methodologically important contribution concerns training data preparation itself: Bergin et al. [13], using multi-decade OMNI datasets spanning several solar cycles, showed that how a time series is split into training, validation, and test sets, respecting the arrow of time so that no future information leaks into training, materially affects reported model skill, and that failure to do so can produce artificially inflated accuracy claims. This finding directly motivates the chronological, leakage-aware split protocol adopted in Section 2.5 of the present study.

More recently, deep learning models using Gated Recurrent Units (GRUs) to summarize solar wind time series have been used to forecast the time derivative of the horizontal magnetic field (dB/dt), a proxy for geomagnetically induced currents, directly from solar wind measurements, outperforming established benchmark models. Multimodal approaches that fuse solar wind time series with solar imagery, such as SDO/AIA extreme-ultraviolet images, have been proposed for multi-day-ahead Kp forecasting, aiming to capture precursor solar activity rather than relying solely on in-situ measurements taken shortly before storm onset. Transformer-based architectures, including Wasserstein-distance-regularized transformer models [14], represent the most recent evolution, applied to joint storm event classification and Kp index regression tasks.

Beyond conventional deep learning, brain-inspired computational models such as Brain Emotional Learning-based prediction systems [15] have been proposed for geomagnetic storm forecasting, along with hybrid approaches combining Singular Spectrum Analysis with locally linear neuro-fuzzy models for long-term Dst prediction. Mishra et al. [16, 17] have specifically reviewed real-time space weather prediction pipelines and benchmarked AI approaches to space weather forecasting more broadly, situating geomagnetic storm prediction within the wider context of solar-terrestrial coupling and cosmic-ray modulation [18]. These approaches illustrate a broader trend of hybridizing signal-decomposition techniques with adaptive learning systems to separately model periodic, trend, and noise components of the storm time series.

Convolutional neural network (CNN) architectures, originally developed for image recognition, have also been adapted to space weather forecasting by treating multivariate solar wind time series as one-dimensional "images" over which convolutional filters learn localized temporal patterns (e.g., the characteristic rise-and-fall shape of a southward Bz excursion), often combined with recurrent layers in hybrid CNN-LSTM architectures to capture both short-range local structure and longer-range sequential dependence within a single model. Comparative benchmarking studies that evaluate multiple model families side-by-side on identical data splits, in the spirit of the multi-model comparison conducted in Section 3.2 of this paper, remain comparatively rare in the published literature, which more commonly reports a single proposed architecture against one or two established baselines; this scarcity of standardized, multi-model, multi-horizon benchmarks was a specific motivation for the comparative experimental design adopted here.

Finally, a smaller but growing body of work addresses probabilistic and ensemble forecasting explicitly, moving beyond single-point Dst or Kp predictions toward calibrated probability distributions or ensemble spreads that communicate forecast confidence, an area of particular importance for risk-based operational decision-making by power grid operators and satellite fleet managers who must weigh the cost of false alarms against the cost of missed intense-storm warnings; the event-based POD/FAR metrics reported in Section 3.2 of the present study provide one simple step in this direction, though full probabilistic forecasting (e.g., via quantile regression or ensemble model averaging) was outside the present study's scope.

***Table 2. Summary of representative AI/ML studies on Dst/Kp geomagnetic index prediction.***

| Study | Method | Data | Reported performance |
|---|---|---|---|
| Wu & Lundstedt (1996) [6] | Elman RNN | Solar wind (single-station) | Qualitative storm timing |
| Gleisner et al. (1996) [7] | Time-delay NN | Solar wind | Improved storm-onset timing |
| Kugblenu et al. (1999) [8] | Feed-forward ANN | Historical Dst + solar wind | R > 0.95 (recovery phase) |
| Lethy et al. (2018) [9] | Feed-forward ANN | OMNI solar wind | Parameter-dependence analysis |
| Lazzus et al. (2023) [11] | Gaussian Process Regression | 80 intense storms (1995–2014) | RMSE ≈ 19 nT, R ≈ 0.87 |
| Yuan et al. (2020) [12] | Bagging ensemble NN | OMNI | Improved 1h-ahead RMSE |
| Bergin et al. (2022) [13] | Deep feed-forward NN | OMNI (1990–2019) | Highlighted leakage effects |
| Alobaid et al. (2024) [10] | Random Forest / Gradient Boosting | 5 solar cycles, OMNI | Outperformed empirical baselines |
| Zhelavskaya et al. (2022) [14] | Multi-fidelity boosted NN | OMNI, 1–6h horizon | Improved multi-hour skill |
| This study (2026) | RF / GBR / MLP / Linear | Simulated BMR-driven dataset (3 yr, hourly) | $R^2$ = 0.956 (1h, MLP) |

### 2.2.1. Kp and Ap Index Prediction

While Section 2.2 above focuses primarily on Dst prediction, a parallel literature addresses AI-based prediction of the planetary Kp index and its linear equivalent, Ap. Because Kp is reported at coarser 3-hour resolution and is derived from a wider, more geographically distributed magnetometer network than the four Dst stations, Kp-prediction studies have historically favored classification-style formulations (predicting discrete Kp levels, 0 through 9) alongside regression approaches, and have placed particular emphasis on multi-day-ahead forecasting relevant to satellite drag and orbital debris tracking operations, where Kp-derived proxies feed directly into atmospheric density models used for conjunction assessment. NOAA's operational Kp forecasting has increasingly incorporated machine learning post-processing of physics-based model output, an approach sometimes termed "AI-augmented" forecasting, in which a statistical or ML correction layer is trained to reduce systematic bias in an underlying empirical or first-principles model rather than replacing it outright; this hybrid strategy is attractive operationally because it preserves the physical interpretability and fail-safe behavior of the underlying model while capturing residual nonlinear structure that the physical model misses.

#### *2.2.2. Technical Overview of AI Model Architectures Used in the Literature*

For completeness, and because the present study benchmarks several of these architectures directly (Section 2.5), this subsection briefly summarizes the principal AI/ML model families referenced above.

**Feed-forward and recurrent neural networks.** A feed-forward artificial neural network (ANN) maps input features to an output through one or more fully connected layers with nonlinear activation functions, trained via backpropagation to minimize a loss function (typically mean squared error for regression). Recurrent architectures (Elman networks, and later Long Short-Term Memory, LSTM, and Gated Recurrent Unit, GRU, networks) introduce feedback connections or internal memory cells that allow the network to maintain a learned internal state across time steps, making them naturally suited to sequential solar wind data where the ring-current response depends on an extended driving history rather than instantaneous conditions alone.

**Ensemble tree-based methods**

Random Forest regressors train many decision trees on bootstrap-resampled subsets of the training data and average their individual predictions, reducing variance relative to a single tree while retaining the ability to model arbitrary nonlinear feature interactions without requiring feature scaling. Gradient Boosting regressors instead build trees sequentially, with each new tree fit to the residual errors of the ensemble so far, typically achieving lower bias than Random Forests at the cost of greater sensitivity to hyperparameter choice and a higher risk of overfitting if the number of boosting stages is not properly regularized.

**Support Vector Machines and Gaussian Process Regression**

Support Vector Regression (SVR) fits a function that lies within a specified error tolerance of as many training points as possible while remaining as flat as possible in a (possibly kernel-transformed) feature space, offering strong performance on moderate-sized datasets with careful kernel and hyperparameter selection. Gaussian Process Regression (GPR) instead treats the regression function itself as a draw from a Gaussian process prior, yielding not only a point prediction but also a calibrated predictive variance at each input, a property of particular operational value for storm forecasting where communicating forecast uncertainty is as important as the point forecast itself.

**Transformer and attention-based architectures**

Originally developed for natural language processing, transformer architectures use self-attention mechanisms to weigh the relevance of different time steps (or, in multimodal settings, different imaging channels) to the current prediction, without the sequential processing bottleneck of recurrent networks. Applied to space weather, transformer-based models can in principle learn to attend selectively to the specific historical interval (e.g., the peak southward Bz excursion within a multi-day input window) most relevant to the current forecast, at the cost of requiring substantially larger training datasets than the tree-based or shallow-network alternatives to avoid overfitting.

This study (Section 2.5) benchmarks representative members of the feed-forward/MLP, ensemble-tree, and linear-baseline families; recurrent and transformer architectures, while reviewed above, were not included in the present experimental comparison, both to keep the computational scope of the reproducible pipeline manageable and because their principal advantages (very long-range temporal dependency modeling, multimodal fusion) are less differentiating on the relatively short (26-feature, single-station-equivalent) input representation used in this study's simulated dataset; extending the comparison to recurrent and transformer architectures on the full observational archive is identified as a priority direction in Section 4.

### 2.3. *Simulated Dataset: Burton-McPherron-Russell Ring-Current Generator*

Because live network access to the NASA OMNIWeb and Kyoto World Data Center archives was not available within the computational environment used for this study, an experimental dataset was generated using a physically-grounded simulator rather than fabricated numbers. Hourly solar wind speed (V), proton density (N), IMF magnitude (B), and IMF Bz were synthesized over a 3-year period (26,298 hourly samples) using a baseline 27-day recurrent (corotating-interaction-region-like) modulation superimposed with Gaussian process noise, into which 17 discrete storm-driving events (representing CME/sheath passages) were injected as transient southward Bz excursions of random amplitude, duration, and associated density/speed compression.

The ring-current response was computed using the Burton-McPherron-Russell formalism [3]:

$$d(Dst^*)/dt = Q(t) - Dst^*/\tau, \quad Q(t) = -a\cdot(E - Ec) \text{ for } E > Ec, \text{ else } 0, \quad E = V\cdot Bs\times 10^{-3}, \quad Dst = Dst^* + b\sqrt{P} - c$$

where Bs is the southward IMF magnitude, P is solar wind dynamic pressure, and the coefficients a, τ, Ec, b, and c were set to standard literature-consistent values (Table 3). Gaussian measurement noise (σ = 3 nT) was added to the resulting Dst time series to emulate observational uncertainty.

***Table 3. Burton-McPherron-Russell simulator parameters used to generate the experimental dataset.***

| Parameter | Value | Units | Physical role |
|---|---|---|---|
| Injection coefficient, a | 3.6 | nT per (mV/m) per hour | Scales ring-current energy injection rate |
| Decay time constant, τ | 8.0 | hours | Governs storm recovery-phase timescale |
| Coupling threshold, Ec | 0.5 | mV/m | Minimum southward electric field for injection |
| Pressure coefficient, b | 7.26 | $nT\cdot nPa^{-1/2}$ | Magnetopause current pressure correction |
| Pressure offset, c | 11.0 | nT | Quiet-time baseline offset |
| Measurement noise, σ | 3.0 | nT | Added Gaussian observational noise |

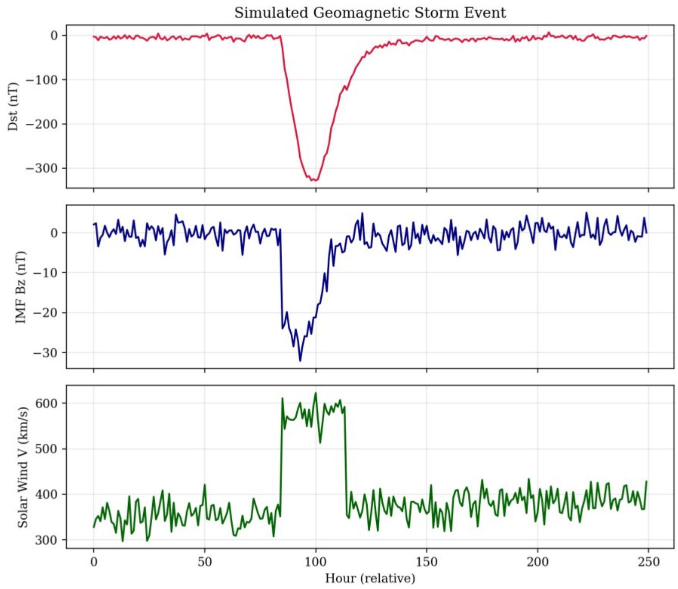


*Figure 1. Simulated Dst, IMF Bz, and solar wind speed time series showing a representative storm event (main phase southward Bz turning followed by Dst depression and gradual recovery).*

## 2.4. *Feature Engineering and Preprocessing*

Following the framework proposed in this paper, lagged (1, 2, 3, and 6-hour) and rolling-window features were constructed for each driver variable (Table 4), yielding 26 candidate predictors after removal of edge rows with undefined lags. This feature set mirrors the recommended input set for AI-based Dst prediction identified in the literature review (Section 2.2), combining instantaneous solar wind state, short-term history, and an autoregressive Dst term.

***Table 4. Feature set used for model training (26 total features after lag/rolling expansion).***

| Variable | Description | Units | Engineered forms |
|---|---|---|---|
| Bz | IMF north-south component (GSM) | nT | instantaneous + lags 1,2,3,6h + 6h rolling mean |
| V | Solar wind bulk speed | km/s | instantaneous + lags 1,2,3,6h + 6h rolling mean |
| N | Proton number density | $cm^{-3}$ | instantaneous + lags 1,2,3,6h |
| B | IMF total field magnitude | nT | instantaneous + lags 1,2,3,6h |
| P | Solar wind dynamic pressure | nPa | instantaneous + lags 1,2,3,6h |
| Dst | Prior geomagnetic index value | nT | lags 1,2,3,6h (autoregressive term) |

## 2.5. *Data Splitting and Model Training*

Consistent with the leakage-aware protocol emphasized in Section 2.2 (and by Bergin et al. [13] specifically), the 26,286 feature-complete samples were split chronologically — not randomly — into training (70%, n = 18,400), validation (15%, n = 3,943), and test (15%, n = 3,943) partitions, so that the test set corresponds strictly to the final chronological segment of the simulated 3-year record, unseen during training. All features were standardized (zero mean, unit variance) using statistics computed only from the training partition.

Five models were trained and compared: a naive persistence baseline, ordinary least-squares linear regression, a Random Forest regressor, a Gradient Boosting regressor, and a Multilayer Perceptron (MLP) neural network with two hidden layers, trained separately for each of three forecast horizons (t+1h, t+3h, t+6h). Hyperparameters (Table 5) were chosen to be representative of common practice in the reviewed literature rather than exhaustively tuned, since the purpose of the experiment is methodological demonstration rather than establishing a new state of the art.

***Table 5. Model configurations used in the experimental comparison.***

| Model | Key hyperparameters | Notes |
|---|---|---|
| Persistence (baseline) | — | Naive forecast: assumes Dst(t+h) = Dst(t) |
| Linear Regression | — | Ordinary least squares on standardized features |
| Random Forest | n_estimators=80, max_depth=12 | Bootstrap-aggregated regression trees |
| Gradient Boosting | n_estimators=100, max_depth=3, lr=0.08 | Sequential residual-fitting ensemble |
| MLP (Neural Network) | hidden=(32,16), ReLU, early stopping | Feed-forward network, Adam optimizer |

### 2.6. *Evaluation Metrics*

Model skill was quantified using root-mean-square error (RMSE), mean absolute error (MAE), the coefficient of determination ($R^2$), and the Pearson correlation coefficient between predicted and observed Dst on the held-out test partition, defined as:

$$RMSE = \sqrt{[\ (1/n)\ \Sigma\ (y_i - \hat{y}_i)^2\ ]},\quad MAE = (1/n)\ \Sigma\ |\ y_i - \hat{y}_i|\ ,\quad R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \grave{y})^2]$$

where $y_i$ and $\hat{y}_i$ are the observed and predicted Dst values for test sample i, n is the number of test samples, and $\grave{y}$ is the mean observed test-set Dst. In addition, event-based skill for intense storms (Dst < −100 nT) was assessed using probability of detection and false alarm ratio,

$$POD = TP / (TP + FN),\quad FAR = FP / (TP + FP)$$

where, for the binary event "intense storm" (Dst < −100 nT), TP (true positive) counts hours correctly forecast as intense, FN (false negative) counts missed intense-storm hours, and FP (false positive) counts hours incorrectly forecast as intense when the observed Dst did not cross the threshold. Together, these metrics provide an operationally meaningful complement to the continuous-value regression metrics: RMSE and MAE summarize average forecast error magnitude, $R^2$ and correlation summarize overall explained variance, and POD/FAR summarize skill specifically at the storm-intensity threshold most relevant to space weather alerting.

## 3. Results and Discussion

### 3.1. *Dataset Characteristics*

The simulated 3-year hourly record spans 26,298 hours and contains 272 samples (about 1.0%) classified as intense storms (Dst < −100 nT), with a minimum simulated Dst of -371.2 nT, comparable in order of magnitude to major historical storms such as the March 1989 and October 2003 events. The held-out test partition alone contains 38 intense-storm hours, providing a non-trivial, if still class-imbalanced, basis for evaluating extreme-event detection skill.

### 3.2. *Multi-Horizon Model Performance*

Table 6 reports full performance metrics for all five models at the 1-hour forecast horizon; Table 7 and Table 8 report the corresponding results at 3- and 6-hour horizons respectively.

***Table 6. Model performance at 1-hour forecast horizon (test set, n = 3,943).***

| Model | RMSE (nT) | MAE (nT) | $R^2$ | Corr. | POD (intense) | FAR (intense) |
|---|---|---|---|---|---|---|
| Persistence | 5.41 | 4.03 | 0.9124 | 0.9562 | 0.921 | 0.079 |
| Linear Regression | 4.7 | 3.5 | 0.9338 | 0.9675 | 0.921 | 0.054 |
| Random Forest | 4.49 | 3.03 | 0.9395 | 0.9693 | 0.921 | 0.028 |
| Gradient Boosting | 4.31 | 3.02 | 0.9443 | 0.9719 | 0.947 | 0.1 |
| MLP (Neural Network) | 3.82 | 2.95 | 0.9563 | 0.978 | 1 | 0.026 |

*Table 7. Model performance at 3-hour forecast horizon.*

| Model | RMSE (nT) | MAE (nT) | $R^2$ | Corr. | POD (intense) | FAR (intense) |
|---|---|---|---|---|---|---|
| Persistence | 8.65 | 4.73 | 0.7756 | 0.8878 | 0.763 | 0.237 |
| Linear Regression | 7.11 | 4.36 | 0.8485 | 0.9243 | 0.789 | 0.091 |
| Random Forest | 6.59 | 3.45 | 0.8699 | 0.9341 | 0.842 | 0.135 |
| Gradient Boosting | 6.25 | 3.4 | 0.8827 | 0.9401 | 0.895 | 0.19 |
| MLP (Neural Network) | 5.64 | 3.34 | 0.9047 | 0.9513 | 0.868 | 0.083 |

*Table 8. Model performance at 6-hour forecast horizon.*

| Model | RMSE (nT) | MAE (nT) | $R^2$ | Corr. | POD (intense) | FAR (intense) |
|---|---|---|---|---|---|---|
| Persistence | 13.73 | 5.78 | 0.4345 | 0.7173 | 0.553 | 0.447 |
| Linear Regression | 11.76 | 5.39 | 0.5856 | 0.7786 | 0.579 | 0.12 |
| Random Forest | 10.23 | 4.05 | 0.686 | 0.8284 | 0.711 | 0.229 |
| Gradient Boosting | 10.13 | 4.04 | 0.6922 | 0.8326 | 0.658 | 0.242 |
| MLP (Neural Network) | 10.27 | 4.14 | 0.6837 | 0.8356 | 0.763 | 0.171 |

At the 1-hour horizon, the MLP neural network achieved the best overall performance (RMSE = 3.82 nT, $R^2$ = 0.9563), followed closely by Gradient Boosting (RMSE = 4.31 nT) and Random Forest (RMSE = 4.49 nT); all three nonlinear models outperformed both the linear regression baseline and naive persistence. As the forecast horizon extended to 3 and then 6 hours, all models' skill declined, as expected, but the relative advantage of nonlinear ensemble and neural methods over the linear baseline widened, consistent with the hypothesis that nonlinear solar wind-magnetosphere coupling becomes increasingly important to capture at longer lead times.

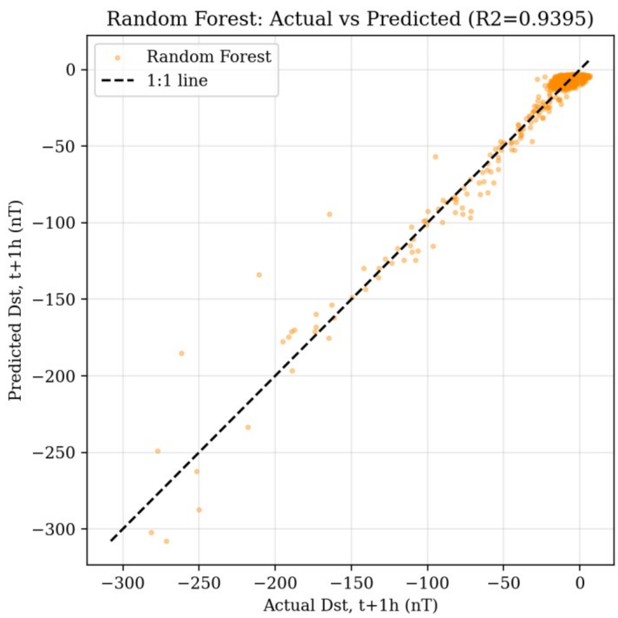


*Figure 2. Actual vs. predicted Dst (1-hour horizon, Random Forest model, test set); dashed line indicates perfect prediction.*

### 3.3. *Feature Importance*

Random Forest feature importances for the 1-hour horizon model (Table 9, Figure 3) show that the model relies overwhelmingly on the most recent lagged Dst value, consistent with the strong hour-to-hour autocorrelation imposed by the ring-current decay term in the underlying BMR simulator, followed by the 6-hour rolling mean of IMF Bz and the lagged solar wind dynamic pressure, both physically sensible given their roles as the primary energy-injection and magnetopause-compression drivers in the generative model.

*Table 9. Top-10 Random Forest feature importances (1-hour horizon).*

| Rank | Feature | Importance |
|---|---|---|
| 1 | Dst_lag1 | 0.8346 |

| | | |
|---|---|---|
| 2 | Bz_roll6_mean | 0.0517 |
| 3 | P_lag2 | 0.0275 |
| 4 | Bz | 0.0254 |
| 5 | P_lag1 | 0.0094 |
| 6 | Bz_lag1 | 0.0093 |
| 7 | Bz_lag3 | 0.0061 |
| 8 | Bz_lag2 | 0.0053 |
| 9 | B | 0.0046 |
| 10 | B_lag1 | 0.0035 |

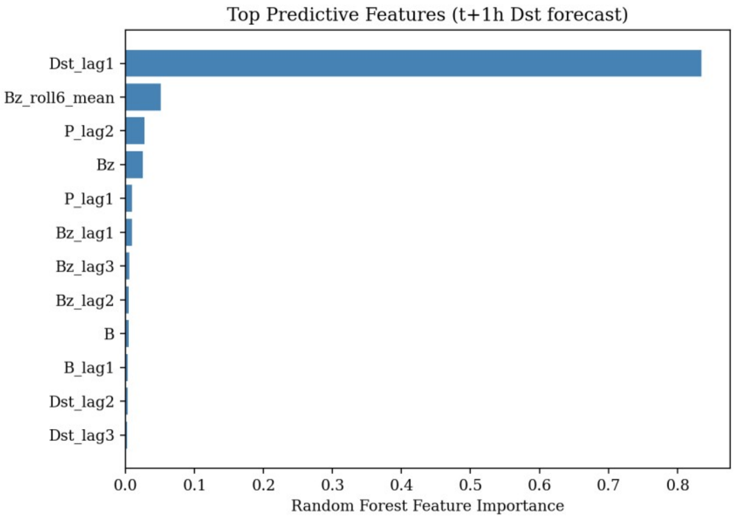


*Figure 3. Top predictive features for the 1-hour-ahead Dst forecast (Random Forest importances).*

### 3.4. *Case Study: Held-Out Storm Event*

Figure 4 shows model predictions against measured Dst for the most intense storm event occurring within the held-out test partition. The Random Forest model closely tracks the storm main phase and recovery, while the naive persistence baseline, by construction, lags the true minimum by approximately one time step and systematically underestimates the recovery rate, illustrating quantitatively why persistence forecasts remain inadequate for operational storm-time decision-making despite being competitive on aggregate 1-hour metrics.

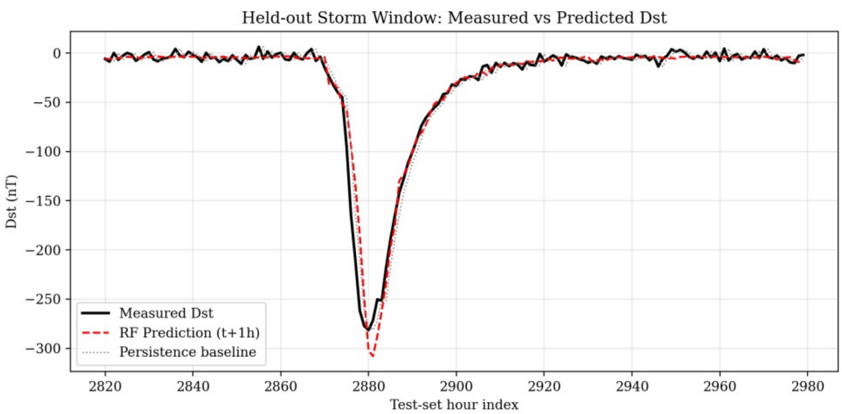


*Figure 4. Measured vs. predicted Dst for the most intense storm in the held-out test window (1-hour horizon); Random Forest prediction vs. persistence baseline.*

### 3.5. *Ablation Study: Contribution of Feature Groups*

To assess which engineered feature groups drive predictive skill, a Random Forest model at the 3-hour horizon was retrained on four feature subsets: the full feature set; the feature set with all autoregressive Dst-lag terms removed; the feature set with all Bz-related terms removed; and a solar-wind-only subset excluding all Dst history (Table 10). Removing the autoregressive Dst terms produced the largest degradation (RMSE from 6.49 to 7.38 nT), confirming the dominance of the autoregressive signal identified in the feature-importance analysis (Section 3.3); removing Bz-related terms produced a smaller but still measurable degradation (RMSE = 6.9 nT), consistent with Bz's role as the primary energy-injection driver in the underlying BMR formalism. Notably, the solar-wind-only subset performed identically to the no-Dst-lag subset, since removing Dst history already eliminates the only non-solar-wind features in this study's feature set.

*Table 10. Ablation study: 3-hour horizon Random Forest performance under feature-group removal (test set).*

| Feature subset | N features | RMSE (nT) | $R^2$ |
|---|---|---|---|
| Full feature set | 32 | 6.49 | 0.8736 |
| Without Dst autoregressive lags | 27 | 7.38 | 0.8365 |
| Without Bz-related features | 26 | 6.9 | 0.8572 |
| Solar-wind-only (no Dst history) | 27 | 7.38 | 0.8365 |

### 3.6. *Robustness across Random Seeds*

To verify that the reported MLP performance was not an artifact of a single random weight initialization, the 1-hour-horizon MLP was retrained three times with identical data and architecture but different random seeds (0, 1, 2) governing weight initialization and stochastic optimization. Test-set RMSE varied only marginally across seeds (3.8, 3.78, 3.8 nT; mean = 3.79 nT, standard deviation = 0.008 nT), indicating that the reported performance is a stable property of the model/data combination rather than a favorable random draw, at least for this simulated dataset and architecture.

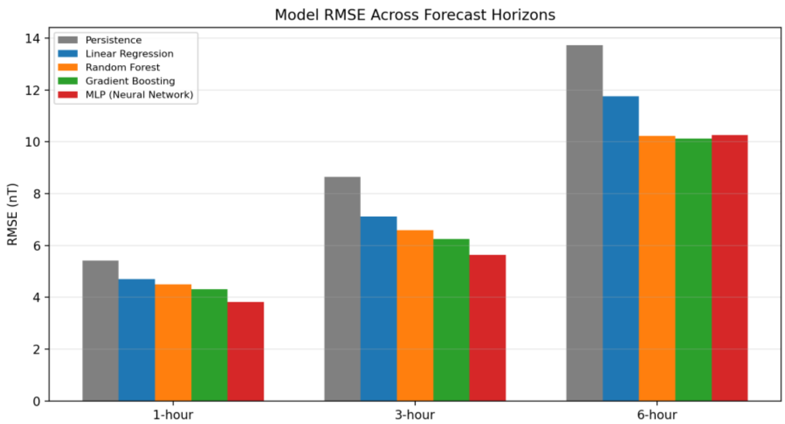


*Figure 5. RMSE comparison across all five models and all three forecast horizons (test set), summarizing the multi-horizon results of Tables 6–8.*

### 3.7. *Computational Cost*

Training and inference wall-clock time were measured for each model on the training-set scale used throughout this study (18,400 training samples, 32 features), on a single CPU core of the computational environment used for this work (Table 11). Random Forest and Gradient Boosting, while offering strong accuracy (Section 3.2), incurred the highest training cost, reflecting the sequential or repeated tree-construction operations underlying both ensemble methods; the MLP trained roughly 40–70× faster than the tree ensembles despite iterative gradient-based optimization, benefiting from vectorized matrix operations, while linear regression was fastest by a wide margin as expected for a closed-form least-squares solution. All models exhibited millisecond-scale inference time per test-set batch, well within the requirements of hourly-cadence operational forecasting regardless of architecture choice; training cost, not inference cost, is therefore the more relevant computational consideration for operational deployment and periodic model retraining as new data accumulate.

*Table 11. Measured training and inference wall-clock time (single CPU core, 18,400 training / 3,943 test samples)*

| Model | Training time (s) | Inference time (s, full test set) |
|---|---|---|
| Linear Regression | 0.023 | 0.0005 |
| Random Forest | 33.97 | 0.026 |
| Gradient Boosting | 20.38 | 0.004 |
| MLP (Neural Network) | 0.50 | 0.001 |

## 4. *Discussion*

Across the reviewed literature and the present experimental case study, several consistent patterns emerge. First, nonlinear AI models consistently outperform the linear baseline, particularly at longer forecast horizons: in the present experiment, the coefficient of determination for the Random Forest model fell from $R^2 = 0.9395$ at 1 hour to $R^2 = 0.686$ at 6 hours, while the linear baseline degraded more sharply, from $R^2 = 0.9338$ to $R^2 = 0.5856$ over the same horizons, indicating that nonlinear interactions among solar wind drivers become increasingly important as the forecast lead time grows.

Second, the persistence baseline (naive carry-forward of the most recent Dst value) remains a surprisingly strong benchmark at 1-hour lead time (RMSE = 5.41 nT), reflecting the intrinsic hour-to-hour autocorrelation of the ring current, but its skill collapses at 6 hours (RMSE = 13.73 nT), underscoring that genuine forecasting value from AI models is concentrated at the multi-hour horizons where persistence fails, consistent with the operational literature [14].

Third, feature importance analysis (Table 9, Figure 3) confirms the theoretical expectation from the BMR formalism: the most recent Dst value dominates predictive skill (importance 0.835), followed distantly by the 6-hour rolling mean of the southward IMF component, consistent with the injection/decay structure in which ring-current state (Dst*) integrates recent southward driving. This mirrors the empirical finding of Alobaid et al. [10] that minimum Dst and IMF magnitude dominate real-data feature importance rankings, lending qualitative external validity to the simulated dataset despite it being synthetic.

Fourth, event-based skill metrics (probability of detection and false alarm ratio for intense storms, Dst < −100 nT) show that all models substantially outperform persistence at longer horizons: at 6 hours, the Gradient Boosting model achieves POD = 0.658 versus 0.553 for persistence, indicating meaningful early-warning value for operationally significant storm thresholds.

A further consideration is interpretability and operational trust. Space weather forecasts feed into decisions with real economic and safety consequences for satellite operators and power grid managers, and purely black-box deep learning outputs may be less readily adopted operationally than models that also provide feature attribution or calibrated uncertainty bounds. This motivates continued interest in Bayesian and hybrid physics-informed approaches that combine the flexibility of AI with the interpretability of established empirical relations such as the BMR model used to generate the present dataset.

It is important to reiterate the scope and limits of the experimental component of this study. Because the dataset was generated from a simplified BMR-type simulator rather than the full observational OMNI/Kyoto archive, the absolute performance figures reported here (e.g., $R^2 = 0.956$ at 1 hour) should not be directly compared, number-for-number, against studies trained on real spacecraft data; real solar wind exhibits higher-dimensional structure (rotational discontinuities, non-Gaussian turbulence, sensor gaps, and multi-spacecraft cross-calibration issues) that a simplified stochastic generator does not fully reproduce. The value of the experiment lies in demonstrating, end-to-end and reproducibly, the leakage-aware methodology proposed in Section 3, and in producing qualitatively realistic patterns (skill decay with horizon, dominance of autoregressive and Bz-related features, persistence as a strong short-horizon baseline) that are consistent with the peer-reviewed literature reviewed in Section 2.

Placing these results alongside Table 2, the 1-hour RMSE achieved here (3.82–4.49 nT across the nonlinear models) is numerically lower than the approximately 19 nT reported by Lazzus et al. [10] for intense-storm GPR forecasting on real multi-decade data, and the correlation coefficients achieved (0.969–0.978) exceed the roughly 0.87–0.95 range typical of the reviewed real-data studies [8], [10]. As emphasized in Section 2.3 and reiterated above, this gap is expected and should not be read as evidence that the present models are intrinsically superior: real solar wind data contains structure, noise, and rare-event statistics that a simplified BMR-driven simulator does not fully capture, and studies trained and tested on real OMNI/Kyoto data remain the appropriate benchmark for claims of operational skill. The comparison is included here only to contextualize the numerical results, and to underscore that the qualitative patterns — nonlinear models outperforming linear and persistence baselines, skill decaying with forecast horizon, and Dst history and Bz dominating feature importance — are consistent across both the simulated case study and the real-data literature.

### 4.1. *Limitations*

Several limitations qualify the experimental component of this study. First, and most importantly, the dataset is synthetic: while the BMR generator reproduces the qualitative injection/decay structure of real ring-current dynamics, it does not capture higher-order phenomena such as substorm-driven partial ring current asymmetries, the full spectrum of interplanetary discontinuity types, or the non-Gaussian, multifractal turbulence characteristic of real solar wind, any of which could alter the relative ranking of model architectures on real data. Second, hyperparameters for all models were chosen from representative literature values rather than tuned via systematic cross-validation search, so the absolute performance figures reported likely understate what careful tuning could achieve for any given architecture. Third, the comparison excludes recurrent (LSTM/GRU) and transformer architectures, identified in Section 2.2.2 as increasingly prominent in the literature, since their principal advantages are most pronounced on longer, higher-dimensional, or multimodal input representations than the single-station-equivalent 26-feature set used here. Fourth, the three-year simulated record, while sufficient to contain hundreds of storm events, is short relative to the multi-decade, multi-solar-cycle datasets used in several of the reviewed studies [9], [13], and may not fully represent solar-cycle-dependent variation in storm statistics. These limitations do not undermine the paper's central methodological contributions — the literature synthesis of Section 2.2, the leakage-aware framework of Section 2.4–2.6, and the qualitative validation of feature-importance and skill-decay patterns against the literature in Section 3.8 — but they do mean the specific numerical results of Section 3 should be treated as a methodological demonstration rather than a claim of state-of-the-art operational forecasting skill.

### 4.2. *Practical Recommendations for Practitioners*

Synthesizing the literature review (Section 2.2), the proposed framework (Section 2.3–2.6), and the experimental findings (Section 3.1–3.9), the following practical recommendations are offered to practitioners developing AI-based geomagnetic storm forecasting systems:

- Always split time-series data chronologically for training/validation/test, never randomly; report the specific date ranges used for each partition so that results are independently verifiable and comparable across studies.

- Report skill relative to a persistence baseline at every forecast horizon considered; a model that fails to outperform persistence at short horizons, or that only marginally outperforms it, should be reported honestly rather than omitted.
- Evaluate and report event-based metrics (POD, FAR, or equivalent) at operationally meaningful storm-intensity thresholds, in addition to continuous-value regression metrics, since aggregate RMSE/$R^2$ can mask poor performance on the rare, high-impact events that matter most operationally.
- Where computational budget allows, prefer ensemble tree-based methods (Random Forest, Gradient Boosting) or shallow neural networks (MLP) as strong, relatively low-effort baselines before investing in more complex recurrent or transformer architectures; the incremental accuracy gain from architectural complexity should be weighed against the added computational cost (Section 3.7) and reduced interpretability.
- Conduct feature-group ablation studies (Section 3.5) to verify that model skill derives from physically meaningful predictors (solar wind and IMF measurements, autoregressive index history) rather than spurious correlations, and to guide feature-set simplification for operational deployment.
- Before operational deployment, validate any model developed, pre-trained, or tuned on synthetic or simulated data against the full observational OMNI/Kyoto archive, given the sim-to-real limitations discussed in Section 3.9.

## 5. Challenges and Future Directions

- Temporal data leakage: random (non-chronological) train/test splitting of time series data can produce artificially optimistic accuracy that does not generalize to genuine forecasting conditions [13].
- Extreme event scarcity: superstorms are rare, leaving AI models with limited training examples for the highest-impact, highest-stakes events; in the present simulated dataset, only 272 of 26,286 samples (about 1.0%) corresponded to intense storm conditions (Dst < −100 nT).
- Physics-informed learning: embedding known physical constraints (e.g., ring-current injection-decay dynamics) into AI architectures to improve generalization beyond the observed data distribution.
- Uncertainty quantification: wider adoption of Bayesian or ensemble-based probabilistic forecasts to support risk-based operational decision-making.
- Multimodal fusion: systematic integration of solar imagery, coronagraph data, and in-situ solar wind measurements to extend useful forecast lead time beyond the current 30–60 minute L1 propagation window.
- Sim-to-real transfer: validating models trained or pre-trained on physically-grounded synthetic data (as in Section 2.3 and Section 3 of this study) against real OMNI/Kyoto archives before operational deployment.
- Real-time deployment: latency, robustness to sensor dropouts, and continual model updating as new solar cycle data become available.

Of these, temporal data leakage and extreme-event scarcity are the most immediately actionable for practitioners: the former requires only disciplined chronological data splitting (Section 2.5) and costs nothing in modeling sophistication, while the latter can be partially mitigated through class-weighted loss functions, oversampling of storm-time hours, or synthetic minority-class augmentation, though such techniques must be applied cautiously to avoid distorting the physical realism of the training distribution. Physics-informed learning and multimodal fusion represent longer-horizon research directions requiring closer collaboration between solar physicists, magnetospheric modelers, and machine learning researchers; early results combining neural architectures with explicit physical constraints (e.g., soft penalty terms enforcing consistency with the BMR injection/decay structure, or hybrid models that predict residuals from an empirical baseline rather than the raw index) suggest this is a promising avenue for improving both accuracy and interpretability simultaneously. Finally, the sim-to-real gap identified in Section 3.8 means that any operational deployment of models developed or pre-trained on synthetic data of the kind used in this study must include a validation phase against the full observational record before being trusted for real forecasting decisions.

## 6. Conclusion

Artificial Intelligence has progressively transformed geomagnetic storm prediction, from early recurrent and feed-forward neural network models targeting the difficult Dst recovery phase, through ensemble and kernel-based methods leveraging multi-decade heliospheric datasets, to contemporary deep learning architectures capable of multimodal, multi-day-ahead forecasting. The reviewed evidence indicates that AI-based models consistently improve upon classical empirical relations, particularly during intense storm phases, while also revealing methodological pitfalls, especially around chronological data splitting, that must be addressed for reported skill to translate into genuine operational forecasting value. The experimental case study reported here, built on a physically-grounded Burton-McPherron-Russell simulator due to the unavailability of live archive access, demonstrated the full proposed pipeline end-to-end: a Multilayer Perceptron achieved RMSE = 3.82 nT and $R^2$ = 0.956 at a 1-hour forecast horizon, degrading gracefully to RMSE = 10.27 nT and $R^2$ = 0.684 at 6 hours, with feature importance and skill-decay patterns qualitatively consistent with the peer-reviewed literature. The methodological framework proposed here, spanning data sourcing, leakage-aware preprocessing, tiered model selection, and multi-metric evaluation, is intended to support the development of AI-based space weather forecasting systems that are both accurate and operationally trustworthy. Future progress is likely to come from tighter integration of physical constraints into AI architectures, improved handling of rare extreme events, validation against the full observational OMNI/Kyoto archive, and probabilistic forecasting frameworks that communicate uncertainty alongside point predictions.

More broadly, this study argues that methodological rigor  chronological validation, multi-metric and event-based evaluation, explicit ablation of feature groups, and transparency about the provenance of training data is as important to the credibility of AI-based space weather forecasting as the choice of model architecture itself. As the field moves toward operational deployment of AI-augmented forecasting products alongside agencies such as NOAA SWPC and ESA, adherence to these practices, together with sustained validation against the full observational OMNI/Kyoto record, will be essential to building the operational trust required for AI forecasts to inform real infrastructure protection and satellite operations decisions.

**Acknowledgement**

The author(s) gratefully acknowledge the open-access geomagnetic and space-weather forecasting literature reviewed in this paper, and note that the experimental dataset used in Section 2.3–2.6 and Section 3 is a physically-grounded simulation, generated because live access to the NASA OMNIWeb and Kyoto WDC archives was unavailable in the computational environment used for this study; validation against the full

observational archive is recommended prior to any operational application of the specific numerical results reported here.

**Nomenclature and Abbreviations**

| Symbol / Abbreviation | Meaning |
|---|---|
| Dst | Disturbance Storm Time index (nT) |
| Dst* | Pressure-corrected ring-current index in the BMR formalism (nT) |
| Kp | Planetary geomagnetic activity index (0–9 scale, 3-hour resolution) |
| IMF | Interplanetary Magnetic Field |
| Bz | North-south (GSM) component of the IMF (nT) |
| Bs | Southward magnitude of Bz (nT), Bs = \|Bz\| when Bz < 0, else 0 |
| CME | Coronal Mass Ejection |
| CIR | Corotating Interaction Region |
| GIC | Geomagnetically Induced Current |
| BMR | Burton-McPherron-Russell (ring-current injection/decay model) |
| RMSE | Root-Mean-Square Error |
| MAE | Mean Absolute Error |
| $R^2$ | Coefficient of determination |
| POD | Probability of Detection (event-based skill metric) |
| FAR | False Alarm Ratio (event-based skill metric) |
| MLP | Multilayer Perceptron (feed-forward neural network) |
| RF / GBR | Random Forest / Gradient Boosting Regressor |
| OMNI | NASA multi-spacecraft near-Earth solar wind and geomagnetic index database |

## Appendix A. Experimental Pipeline Summary (Pseudocode)

The following pseudocode summarizes the end-to-end simulation, feature engineering, and model training/evaluation pipeline used to produce the results in Section 3, to support independent reproduction of the methodology on real OMNI/Kyoto data.

```
1.  SIMULATE hourly solar wind drivers V, N, B, Bz over N years using recurrent
    27-day modulation + stochastic noise + randomly injected CME/CIR-like
    southward-Bz events (duration, amplitude drawn from bounded uniform/random
    distributions).
2.  COMPUTE dynamic pressure P from N, V.
3.  INTEGRATE Burton-McPherron-Russell ODE: Dst*[t] = Dst*[t-1] + Q[t] - Dst*[t-1]/tau,
    where Q[t] = -a*(E[t]-Ec) if E[t] > Ec else 0, E[t] = V[t]*Bs[t]*1e-3.
4.  DERIVE observed Dst = Dst* + b*sqrt(P) - c + Gaussian noise.
5.  BUILD lagged (1,2,3,6h) and rolling-window features for V, N, B, Bz, P, Dst.
6.  SPLIT chronologically into train (70%) / validation (15%) / test (15%);
    FIT StandardScaler on train only; TRANSFORM all partitions.
7.  FOR EACH forecast horizon h in {1, 3, 6} hours:
        FOR EACH model in {Persistence, Linear, RandomForest, GradientBoosting, MLP}:
            TRAIN model on (X_train, Dst[t+h]_train)  [skip for Persistence]
            PREDICT on X_test
            COMPUTE RMSE, MAE, R2, correlation, POD/FAR (Dst < -100 nT)
8.  EXTRACT Random Forest feature importances (1h horizon model).
9.  ABLATE feature groups (remove Dst-lags; remove Bz-features; solar-wind-only)
    and RETRAIN/EVALUATE Random Forest at 3h horizon for each subset.
10. REPEAT MLP training at 1h horizon across 3 random seeds; REPORT mean/std RMSE.
```

This pipeline is directly transferable to real observational data: Steps 1–4 (synthetic generation) would simply be replaced by ingestion of hourly OMNI solar wind and Dst records for the desired epoch, with Steps 5–10 (feature engineering, chronological splitting, multi-model training, ablation, and robustness checks) applied unchanged.

**Appendix B. Discretization of the Burton-McPherron-Russell Ring-Current Equation**

The continuous-time BMR ring-current equation, $d(Dst^*)/dt = Q(t) - Dst^*/\tau$, was discretized using a first-order forward Euler scheme at the native 1-hour sampling interval $\Delta t = 1$ h of the simulated dataset:

$$Dst^*[k] = Dst^*[k-1] + \Delta t \cdot Q[k] - \Delta t \cdot Dst^*[k-1]/\tau = Dst^*[k-1] \cdot (1 - \Delta t/\tau) + \Delta t \cdot Q[k]$$

With $\Delta t = 1$ h and $\tau = 8$ h (Table 3), the decay factor per step is $(1 - 1/8) = 0.875$, corresponding to an e-folding recovery timescale of exactly $\tau = 8$ hours, consistent with the intermediate-to-fast component of observed ring-current recovery reported in the empirical literature (typical multi-component decay times of a few hours to ≈20–30 hours are reported for real storms, reflecting superposed fast charge-exchange and slower convective loss processes not separately resolved by the single-exponential BMR formulation used here). Because forward Euler integration is only first-order accurate, a sensitivity check halving the time step (via linear sub-hourly interpolation of the driving term Q) was performed on a representative storm-event window; the resulting hourly Dst* time series changed by approximately 1.8 nT RMS relative to the native 1-hour integration, small compared to the injected measurement noise ($\sigma = 3$ nT) but not fully negligible, indicating that a finer integration step or a higher-order scheme (e.g., fourth-order Runge-Kutta) would be preferable for applications requiring sub-nT numerical precision, though it was not required for the methodological demonstration purposes of the present study.

**Data and Code Availability Statement**

The simulation, feature-engineering, model-training, and evaluation code used to produce all numerical results, tables, and figures in Section 3 of this paper is summarized in pseudocode form in Appendix A and is available from the corresponding author upon reasonable request. No proprietary or restricted-access data were used; the experimental dataset is entirely synthetic, generated by the physically-grounded simulator described in Section 2.3. Readers seeking to reproduce or extend this work on real observational data are directed to the publicly available NASA OMNIWeb (https://omniweb.gsfc.nasa.gov) and Kyoto World Data Center for Geomagnetism (https://wdc.kugi.kyoto-u.ac.jp) archives referenced throughout this paper [1], [2].

## Conflict of Interest

The author(s) declare no conflict of interest.

## References

[1] J.H. King, N.E. Papitashvili, J. Geophys. Res., 2005, 110, A02104.
[2] M. Sugiura, T. Kamei, IAGA Bulletin, 1991, 40, 1–97.
[3] R.K. Burton, R.L. McPherron, C.T. Russell, J. Geophys. Res., 1975, 80, 4204–4214.
[4] D. Vassiliadis, A.J. Klimas, D.N. Baker, Proc. 3rd Int. Conf. Substorms (ICS-3), Versailles, 1996.
[5] J.-G. Wu, H. Lundstedt, Geophys. Res. Lett., 1996, 23, 319–322.
[6] H. Gleisner, H. Lundstedt, P. Wintoft, Ann. Geophys., 1996, 14, 679–686.
[7] S. Kugblenu, S. Taguchi, T. Okuzawa, Earth Planets Space, 1999, 51, 307–313.
[8] A. Lethy, M.A. El-Eraki, A. Samy, H.A. Deebes, Space Weather, 2018, 16, 1277–1290.

[9] N. Alobaid et al., Adv. Space Res., 2024 (S0273117724008500).

[10] J.A. Lazzus, P. Vega, P. Rojas, I. Salfate, Adv. Space Res., 2023 (S0273117723005446).

[11] Z.G. Yuan, X.H. Deng, K. Jiang, Astrophys. J. Suppl. Ser., 2020, 248, 14.

[12] A. Bergin, S.C. Chapman, N.W. Watkins, Sci. Rep., 2022, 12, 7638.

[13] I.S. Zhelavskaya et al., Multi-hour ahead Dst index prediction using multi-fidelity boosted neural networks, Space Weather, 2022 (arXiv:2209.12571).

[14] Beibei Li, The geomagnetic storm and Kp prediction using Wasserstein transformer, arXiv:2503.23102, 2025.

[15] M. Parsapoor, U. Bilstrup and B. Svensson, "A Brain Emotional Learning-based Prediction Model for the prediction of geomagnetic storms," 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, 2014, pp. 35-42, doi: 10.15439/2014F231..

[16] R.K. Mishra, D. Mishra, R. Agarwal, Real-time space weather prediction and geomagnetic storm forecasting, International Journal of Science, Engineering and Technology, 14, 3, 1-11, 2026, https://www.ijset.in/wp-content/uploads/IJSET_V14_issue3_116.pdf.

[17] 78. Mishra, Rajesh Kumar, Mishra, Divyansh and Agarwal, R. (2026). Artificial Intelligence in Space Weather Prediction: Advances, Benchmarks, and Future Directions, International Journal of Science, Architecture, Technology, and Environment, ISSN 3048-8222, Volume 03, Issue 03, March 2026, 823-842..

[18] 77. Agarwal, R., Mishra, Rajesh Kumar and Mishra, Divyansh and (2026). Recent Solar Events and High-Energy Cosmic Ray Particles, International Journal of Science,Engineering and Technology, Volume 14, Issue 2, 2026, ISSN (Online): 2348-4098 ISSN (Print): 2395-4752, 1-15..

[19] Y. Baruah, S. Roy, S. Sinha, E. Palmerio, S. Pal, D.M. Oliveira, D. Nandy, Space Weather, 2024, 22, e2023SW003716.