Research on Forecasting Geomagnetic Storm Based on Self-Attention Mechanism

Jiacheng Li*

Abstract—**Real-time prediction of geomagnetic storms has become increasingly important with the rapid development of science, technology, and space exploration. While data-driven models provide accuracy and flexibility, they often struggle with noise, missing values, and modeling long-term dependencies. These limitations hinder their ability to accurately predict extreme events. Physical models provided strong theoretical frameworks to explain geomagnetic storm mechanisms. However, their reliance on complex parameters and precise observational data limited their adaptability to dynamic conditions. To address these issues, this study proposes a DTW-Attention model based on the Self-Attention mechanism and the Dynamic Time Warping (DTW) method for geomagnetic storm prediction. The model uses an embedding layer to project time series data into a high-dimensional space. A multi-layer encoder captures both short-term and long-term dependencies. Positional encoding enhances the model's temporal sensitivity. During optimization, the DTW-Attention model improves time-series alignment. Experimental results show that the proposed model significantly improves prediction accuracy compared to the classic deep learning methods. The DTW-Attention model combines the temporal alignment of DTW with the global modelling capabilities of the Self-Attention mechanism, significantly reducing short-term and long-term errors. The model further exhibits improved stability and robustness across both medium-term and long-term forecasting horizons. Multi-line time series plots further confirm the model's effectiveness in capturing short-term trends and long-term volatility.**

Index Terms—**Self-Attention Mechanism, Geomagnetic Storm, Dynamic Time Warping, Disturbance Storm Time**

I. Introduction

Geomagnetic storms are intense space weather phenomena triggered by solar activity. Solar wind and Coronal Mass Ejections (CME) interact with earth's magnetic field, generating storms. Fluctuations in Earth's magnetic field can severely disrupt spacecraft, communication infrastructure, and power grids [1-2]. Accurately predicting the timing, intensity, and spatial extent of geomagnetic storms is essential for implementing proactive measures to mitigate their impact. Despite recent advances, precise geomagnetic storm prediction remains challenging, particularly for applications in satellite, aviation, and communication systems.

Physical and data-driven models form the basis of current prediction methods. Physical models primarily rely on observational data such as solar wind speed, temperature, density, and solar activity. By incorporating these data into magnetosphere models, researchers aim to predict geomagnetic storms more accurately. Physical models explain storm formation based on strong theoretical principles but face several limitations. For example, these models involve numerous complex parameters, and accurate computation relies on high-quality, real-time observational data [3]. In contrast, data-driven models, such as machine learning and deep learning, are trained on big historical datasets to recognize potential geomagnetic storm patterns [4-5]. While data-driven models enhance prediction accuracy and adaptability, they face challenges like noise, missing values, and insufficient high-quality data. Additionally, these models struggle with rare or previously unseen storm events and generally lack the explanatory power of physical models [6]. While deep learning can effectively solve complex problems, its applications for real-time predictions are still constrained by the necessity for intensive training, extensive hyperparameter tuning, and high computational costs [7-9].

Geomagnetic storm prediction is inherently a time-series problem, as its occurrence and progression depend heavily on historical observations. The advancement of machine learning, particularly deep learning models, has significantly improved geomagnetic storm prediction time-series models, such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, have demonstrated superior performance in geomagnetic storm prediction. For instance, Priatna et al. [10] applied RNN and LSTM models in 2020 to analyze prediction duration and compare various forecasting periods, demonstrating that shorter periods result in higher accuracy. The study highlighted the advantages of RNN and LSTM models in short-term weather prediction but also identified challenges such as overfitting and low training efficiency when processing complex geomagnetic storm data. Cristoforetti et al. [11] utilized solar wind and interplanetary magnetic field data as inputs for a Deep Neural Network (DNN) to forecast the Disturbance Storm Time (Dst) index during geomagnetic storms. This indicates that DNN may adapt to various space weather conditions, ranging from quiet periods to severe geomagnetic storms. The DNN model has several notable limitations, including low interpret-ability, strong dependence on data quantity and quality, and increased computational and training demands as network depth increases. In 2023, Uyanık et al. [12] utilized image processing techniques to extract spatial-temporal correlations. They constructed the Time Evolutionary Correlation (TEC) images using a time frequency representation and input them into a Convolutional Neural Network (CNN). The study reported

Manuscript received March 1, 2025; revised August 7, 2025.

Jiacheng Li is a postgraduate student of School of Mapping and Geographical Science, Liaoning Technical University, Fuxin, Liaoning, CO 123000 China (corresponding author to provide phone: +086-18342867258; email:18342867258@163.com).

that the approach achieved an accuracy of 89.31% in predicting geomagnetic storms. A key advantage of CNN is its ability to integrate temporal and spatial information. However, they also present significant limitations, such as high computational costs, strong data dependency, and ability of limited explanation.

In recent years, time-series prediction has increasingly adopted the Self-Attention mechanism, which effectively captures long-range dependencies in sequences [13]. The Self-Attention mechanism offers the advantage of dynamically assigning weights to different sequence segments while incorporating global information, free from the constraints of traditional models' local window sizes. This capability allows the Self-Attention to efficiently capture dependencies in time-series data, such as geomagnetic storms. The Self-Attention mechanism dynamically adjusts attention to different time steps based on past solar activity and geomagnetic field variations, enhancing the accuracy of geomagnetic storm predictions [14]. Compared to the traditional RNN or LSTM based models, the Self-Attention mechanism more effectively processes data with long time spans while significantly enhancing real-time performance and computational efficiency due to its strong parallel computing capabilities. The development of geomagnetic storms is influenced by solar activity occurring over multiple time scales. Traditional models often fail to capture inter-hourly and inter-day dependencies because of limited receptive fields and vanishing gradient problems. In contrast, the Self-Attention mechanism connects arbitrary time steps through a global attention matrix, allowing more accurate modeling of both the onset and recovery phases of geomagnetic storms.

Despite the application of deep learning models such as LSTM and CNN in geomagnetic storm prediction, two major challenges persist: error accumulation in long-term forecasting, and the susceptibility of DTW to noise and its limited flexibility in temporal alignment. This study presents a geomagnetic storm prediction DTW-Attention model. The new model combines the temporal alignment of DTW with the global modelling capabilities of the Self-Attention mechanism. The model effectively captures long-range dependencies in time-series data, integrates global information, and enhances both prediction accuracy and real-time performance. Experimental results demonstrate that the proposed model outperforms classic deep learning methods across multiple key metrics. Furthermore, this study refines the model's training strategy to align with the characteristics of geomagnetic storm data, improving robustness and generalization. These advancements contribute to the theoretical foundation and technical development of real-time space weather warning systems.

II. Related Work

Deep learning extensively relies on the Self-Attention mechanism, a fundamental concept essential for data processing and analysis. The Self-Attention mechanism captures dependencies between input elements by dynamically assigning attention weights, enabling the model to focus on salient information. This mechanism enables the model to focus on the most relevant information for the prediction task. This mechanism is crucial because it enables the model to concentrate on the information that is the most beneficial for the specific task, thus enhancing its performance and accuracy [15]. One of the significant advantages of the Self-Attention mechanism is that it significantly increases computational efficiency. This efficiency is achieved through parallel processing of input data, which means that data is processed simultaneously, making the computational process faster. Moreover, it excels in efficiently capturing long range relationships that exist in sequential data. In the Self-Attention mechanism, each individual input element is meticulously mapped to a query, a key, and a value. Subsequently, the model assigns weights to these values. The determination of the correlation between the query and the key is a critical phase in this process, as it enables the model to assign importance to the different data points based on their relevance to the task. This weighting mechanism ensures that the model effectively prioritizes relevant data points, ultimately improving the overall performance of the model enabled by this mechanism.

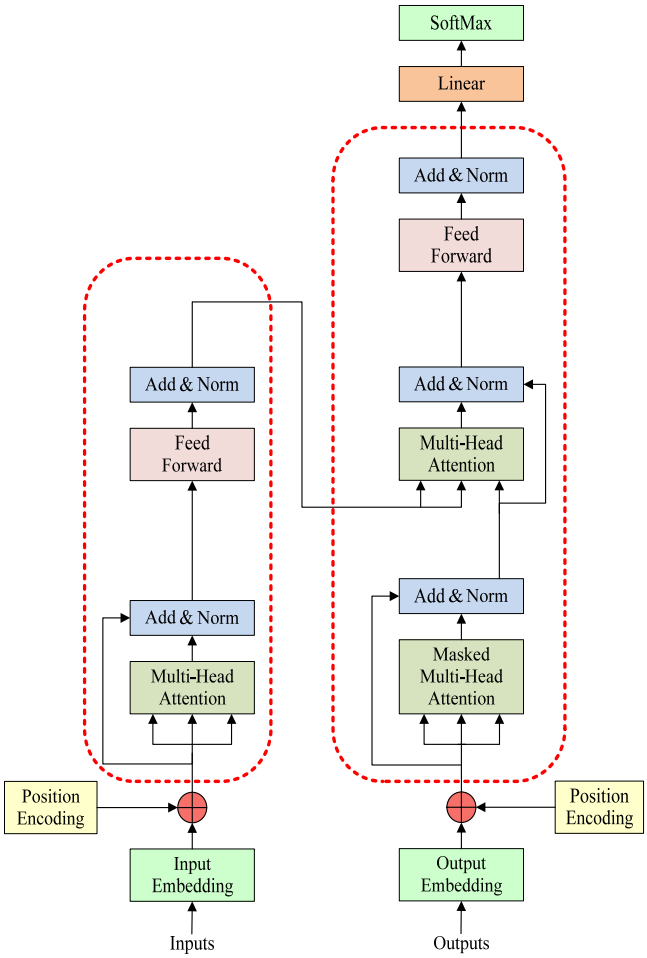


Fig.1. Transformer model architecture diagram

Transformer model architecture [16], illustrated in Fig. 1, which incorporates several key technical features, including positional encoding, the Self-Attention mechanism, encoder and decoder structure, and multi-head attention. The integration of these features enables the Transformer model to capture contextual information and long-range dependencies better when processing sequence data, thereby improving the model's performance.

Transformer model is a deep learning architecture based on the Self-Attention mechanism, designed to handle large amounts of input data simultaneously and model connections over time. Its primary function is to quickly capture global information in sequential data using multiple Self-Attention layers and positional encoding, avoiding the gradient vanishing issue encountered in the traditional RNN with long sequences. In our research, we introduced a lightweight Transformer architecture based on a time attention mechanism. Weighted summation of input sequences using learnable time weights to highlight historical moments that have the greatest impact on the prediction target.

A. Input Mapping

First, we map the input data, which consists of words in a sequence, into three distinct linear transformations: query, key, and value. Suppose the input is a sequence of vectors $X = \{x_1, x_2, \ldots, x_T\}$; each x_i is a high-dimensional vector. The weight matrix is obtained by learning maps X to query, key, and value, respectively. We receive an input sequence $X \in \mathbb{R}^{T \times d_{\text{model}}}$. The query ($Q$), the key ($K$), and the value ($V$) are derived through linear transformations using matrices $Q = XW^Q$, $K = XW^K$, and $V = XW^V$, where W^Q, W^K, and W^V are the weight matrices acquired from training. Dimensions are $d_{\text{model}} \times d_{\text{key}}$ and $d_{\text{model}} \times d_{\text{value}}$. Fig. 2 displays the input mapping.

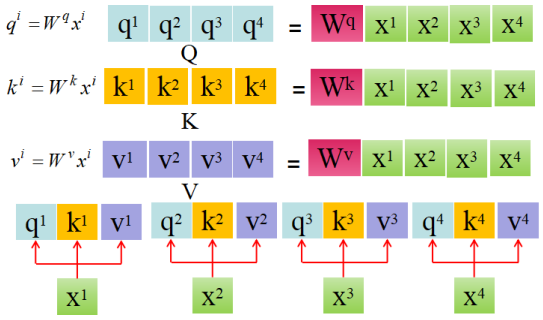


Fig.2. Input mapping

B. Data Processing

1) Calculate scores: The query vectors and the key vectors are computed by taking the dot product of the similarity, i.e., the correlation between each query vector and all other inputs. $Scores(Q, K)$ is calculated as shown in Equation (1).

$$Scores(Q, K) = \frac{QK^T}{\sqrt{d_{key}}} \tag{1}$$

The matrices Q and K represent queries and keys, respectively, whereas d_k represents the key's dimension. $\sqrt{d_{key}}$ is the scaling factor used to normalize the dot product, preventing excessively large values.

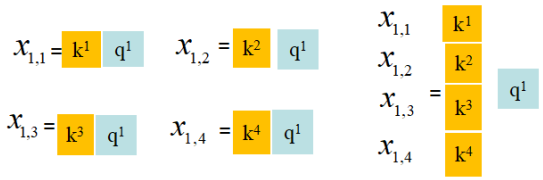


Fig.3. Processing of dot product

The correlation between every two input vectors is calculated using the obtained Q and K, that is, the value of attentions X. X is calculated in various ways, usually by dot product. The dot product is shown in Fig. 3.

2) Calculate weights: To determine the relative importance of one element over others, it calculates the similarity of each query to every key. The dot product, as shown in Equation (2), is the standard way to compute similarity. The result of this dot product is normalized by a Softmax function to ensure that all weights sum to 1.

$$Weights = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{key}}}\right) \tag{2}$$

3) Sum weights: The final output vector is created by weighting and adding the values (V) once the weights have been obtained. This procedure is demonstrated by Equation (3).

$$Output = Weights \times V \tag{3}$$

Each output vector contains weighting information for all input elements, with weights determined by the similarity between the query and the key. The model generates a new set of representations by adjusting each input element based on weighting information from preceding items. The network forwards this output sequence to the next layer for further processing. These principles form the core of the proposed model. Section III provides detailed information on its implementation.

III. Proposed Method

A. Model Composition

We define a DTW-Attention model for tackling the prediction task of time-series data on geomagnetic storms. The traditional DTW alignment method [17] assesses the temporal shift between two time-series. The new method calculates the similarity between two time-series and computes the matching cost based on the minimum distance path. The Self-Attention computes the embedded representation of two time-series through the Transformer layer and measures the similarity of the two-series through the Euclidean distance. The model first turns the input time-series into a high-dimensional space using an embedding layer. It then encodes the data using a Self-Attention process to create a new representation. The similarity matrix of the embedding vectors is subsequently calculated, and time-series alignment is performed by reducing the cost. We employed the Transformer encoder architecture and incorporated positional encoding to enhance the model's sensitivity to time-series data. The model comprises a series of stacked Self-Attention layers that constitute the encoder. The proposed architecture consists of the following core components:

1) Embedding Layer: We convert the input time-series data into a fixed dimensional vector representation. The embedding layer maps the input geomagnetic storm time-series to a high-dimensional space. The embedding layer projects the input into a high-dimensional space via linear transformation in Equation (4).

$$E = XW^E,\ W^E \in \mathbb{R}^{1 \times d_{\text{model}}} \tag{4}$$

A grid search is used for validation, and $d_{\text{model}} = 64$ is selected to balance model capacity and overfitting, preventing an increase in validation errors. The high-dimensional embedding isolates noise from valid signals and captures nonlinear time-series features, including sudden phase shifts and recovery trends in geomagnetic storms. In each time step, this layer maps the input features to a built-in vector space. Through this mapping, the model can capture more feature information and improve its representation.

2) Positional encoding: To preserve sequential information in time-series data, we implement positional encoding for each time step. Strengthening the localization constraint of the Self-Attention mechanism helps the model better understand the sequential arrangement of elements within time-series.

The position encoding employs a learnable parameter $P \in \mathbb{R}^{1\times T\times d_{\text{model}}}$ rather than a fixed sinusoidal function. This design dynamically updates position weights during training, enhancing adaptability to the non-stationary characteristics of magnetic storm data.

3) Transformer encoder: The transformer encoder is the model's core component. It comprises multiple layers of Self-Attention and feed-forward networks that determine how each sequence time step is connected to the others. Each Self-Attention layer concurrently assesses the similarity of all sequence positions. The Transformer encoder comprises four stacked Self-Attention layers, each with four attention heads. The multi-head mechanism enables the model to simultaneously capture features at different time scales, including short-term perturbations and long-term trends. The feed-forward network has a dimension of 256, and the activation function is used to enhance nonlinear modeling. The vanishing gradient problem is alleviated using residual connections and layer normalization. It subsequently assigns weights and aggregates the information to model global dependencies. Following multiple coding layers, Equation (5) represents the sequence as Z_{out}.

$$Z_{out} = Encoder(X) \tag{5}$$

4) Output layer: The model maps the high dimensional representation of the encoder output to a scalar prediction through a fully connected layer. The output of this layer is the model's prediction, representing the intensity of a geomagnetic storm at a future time. The output layer transforms the encoder output into scalar predictions via a fully connected layer, as shown in Equation (6).

$$\hat{y}_t = Z_{out}W^o + b^o,\ W^o \in \mathbb{R}^{d_{\text{model}}\times 1} \tag{6}$$

Where b^o is the bias term for the output layer; W^o is the weight matrix of the output layer with dimension $d_{\text{model}} \times 1$. $\hat{y}_t$ represents the predicted value of the model for the *i*-th time step.

The embedding layer and positional coding enable the model to transform the input data into a high-dimensional embedding space. The Self-Attention layer of the transformer then processes the data, effectively leveraging the long-range dependencies in the time-series to address the complex temporal issues associated with geomagnetic storm data. The model finally outputs the predicted values.

B. Evaluation Metrics

We use the theoretical framework from Section II to create a new DTW-Attention metrics for testing geomagnetic storm models. The goal is to make predictions more accurate in time-series data analysis. The goal of the DTW-Attention metrics is to find the path of correspondence between $T = \{t_1, t_2, ..., t_T\}$ and $P = \{p_1, p_2, ..., p_p\}$. The model assesses their similarity by calculating their Self-Attention representations. The embedding representations E_T and E_P are derived from Self-Attention modeling, as illustrated in Equations (7) and (8).

$$E_T = Attention(T) \tag{7}$$

$$E_p = Attention(P) \tag{8}$$

Where E_T and E_P are embedded representations of the time-series *T* and *P*, respectively.

The above equation encompasses the following calculations:

1) Compute similarity matrix: The similarity matrix $\boldsymbol{S}$ is obtained by calculating the Euclidean distance between the embedding vectors of the time-series *T* and *P*, as outlined in Equation (9). Unlike conventional DTW, which directly computes the distance between original sequences, DTW-Attention constructs a similarity matrix in the embedding space. The Transformer encoder generates the embedding representations E_T and E_P, which suppress noise and capture semantic features to align the onset of the magnetic storm.

$$\mathrm{S}_{ij} = \left\| E_T^{(i)} - E_P^{(j)} \right\|_2 \tag{9}$$

$\|\cdot\|_2$ represents the Euclidean distance. In the time-series *T* and *P*, the time step indices are represented by *i* and *j*.

2) Compute least cost path: The least cost path is determined by identifying the shortest route in the similarity matrix, and this path is derived using dynamic programming methods. Assume the path is designated as $\pi = \{(i_1, j_1), (i_2, j_2), ..., (i_k, j_k)\}$.

The DTW cost represents the cumulative distance of these path points, as illustrated in Equation (10).

$$\mathrm{DTW}(T,P) = \sum_{(i_j,j_j)\in\pi} S_{i_j,j_j} \tag{10}$$

Where S_{i_j,j_j} represents the similarity associated with the path points. The minimum-cost path is computed using dynamic programming, subject to the constraints of monotonic and continuity. DTW-Attention improves upon conventional methods by dynamically adjusting path weights and enabling multi-scale alignment. Attention weights regulate the local sensitivity of aligned paths, facilitating a more flexible alignment during the magnetic storm recovery phase. Additionally, the multi-head mechanism constructs multiple similarity matrices, and the final path is derived by averaging these results, enhancing robustness. By calculating DTW cost and the path between the predicted and actual values of the model, we can measure its predictive performance.

Assuming that the predicted sequence is $\hat{P}$ and the true

sequence is P , DTW-Attention metrics is presented in Equation (11).

$$\hat{S}_{i,j} = \frac{S_{i,j}}{\sum_{i=1}^{P} S_{i,k}}, \forall i \in \{1,2,...,T\}, j \in \{1,2,...,P\} \quad (11)$$

In Equation 11, T and P represent the lengths of the reference time series T and the target time series P, i.e., the number of time steps, respectively. $S_{i,j}$ is the Euclidean distance between step i of the reference sequence and step j of the target sequence. $\hat{S}_{i,j}$ represents the normalized similarity weight that indicates the relative importance of step i in relation to step j in the path alignment. The similarity matrix, S, along with the normalized matrix, $\hat{S}$, ensures that the sum of each row equals 1. The computation of the normalization matrix $\hat{S}$ converts the similarity matrix into a probability distribution to avoid bias due to sequence length differences.

The DTW cost is shown in Equation (12) and is used to measure the alignment error of the predicted sequence $\hat{P}$ with the true sequence P .

$$DTW(T,P) = \sum_{(i,j)\in\pi} S_{i,j} \quad (12)$$

IV. Experiments

A. Dataset and Preprocessing

The study utilizes geomagnetic storm datasets from the OMNI database [18]. This database has been systematically collecting data on various aspects of the space environment near earth since 1963, including the solar wind and geomagnetic field. Specifically, the features examined in this study were measured between 00:00 on 14 January 2001 and 23:00 on 31 December 2016, resulting in a total of 139,944 entries in the full extracted dataset. The test set includes data from the OMNI database for July and December of each year, featuring geomagnetic storm cycle characteristics and time frames. In contrast, the training set consists of data from other months. The data is then divided chronologically, and a custom function randomly selects a certain percentage of training samples to create the validation set. The dataset is split into training, validation, and test sets for model training, testing, and evaluation. The proposed model is programmed using Python and the TensorFlow 2.18.0 machine learning library, and is trained and tested on a 64-bit Windows 10 operating system.

The performance of geomagnetic storm predicted models critically depends on the quality of data preprocessing [19]. Preprocessing includes loading and cleaning data, generating features, converting timestamps, handling outliers, and imputing missing values. We apply a comprehensive set of preprocessing steps to ensure data quality, establishing a reliable foundation for subsequent model training.

The preprocessing is performed to customize it to the requirements of the deep learning model. Specific steps include:

1) Noise reduction and standardization: To maintain clean data and reduce noise in the model input, we first remove missing values. Thereafter, standardization converts features into a distribution with mean 0 and variance 1. This ensures consistent feature scale, prevents features with large magnitude differences from influencing model training, and accelerates convergence while enhancing predicted accuracy. Equation (13) illustrates the normalization transformation for each feature y_i .

$$\hat{x}_i = \frac{x_i - \mu}{\sigma} \quad (13)$$

Where μ represents the mean of the feature and σ denotes the standard deviation. x_i and $\hat{x}_i$ represent the original and normalized eigenvalues, respectively.

2) Feature extraction and target variable: Feature extraction and target variable identification are performed on the data, with the target variables being time lagged to facilitate prediction. Equation (14) illustrates the lag equation for the target variable, where $X = \{x_1, x_2, ..., x_n\}$ represents the feature dataset and $Y = \{y_1, y_2, ..., y_n\}$ denotes the target variable. We use Y' for time-series forecasting.

$$Y' = \{y_2, y_3 ..., y_n\} \quad (14)$$

The OMNI dataset contains many features, but not all significantly influence predicted outcomes. Therefore, during data preprocessing, we select and extract relevant features to minimize overfitting, streamline computational complexity, and enhance model performance. Feature engineering involves deriving input features and target variables from raw data to address time-series problems. Some feature data, such as historical data from various time points, is used as input features in this paper.

Some of these features are the average field magnitude |*B*|, the proton density, the IMF z-component Bz, the plasma flow speed, and the geomagnetic Dst index. We extract these features from the OMNI database. The Dst index[20-21] typically indicates the strength of a geomagnetic storm; a lower Dst index signifies a stronger storm. Therefore, we focus on predicting the future Dst index. This process reduces data dimension, alleviates the computational burden during training, and decreases the risk of overfitting caused by high-dimensional data. Additionally, new features are generated from the original data, specifically the increments of the target variable Dst. Incremental features help the model capture trend variations in time-series data.

The dataset's year, day, hour, and other relevant details are converted into a precise timestamp format for time-series analysis and future queries. The input data were structured as 3D tensors with *X* time steps. For each sample, we use historical data from the previous *time_back* points to forecast the target variable for the subsequent *time_forward* points. Designate the input data as $X = \{x_1, x_2, ...x_T\}$, with T being the total number of samples. Each sample x_i represents a sequence of preceding time intervals $X_i = \{x_{i-time\_back}, x_{i-time\_back+1}, ..., x_{i-1}\}$. The corresponding data y_i for the next *time_forward* step is provided in the output $Y_i = \{y_i, y_{i+1}, ..., y_{i+time\_forward-1}\}$.

3) Storm date extraction and matching: External storm date data was loaded and aligned with the time-series data to identify storm events in the test set. The processed training set, validation set, and test set are stored as HDF5 format

files. By preprocessing the data, as previously mentioned, the dataset only retains the features relevant to geomagnetic storm prediction, reducing computation and increasing model training efficiency. This also aids the model in capturing the trending changes in the time-series data.

B. Model Trainning

To train the model efficiently, we structured the data to facilitate effective learning of long-term dependencies in time-series signals. The dataset enables forecasting at multiple time horizons, including 1-hour, 3-hour, and 6-hour targets, which enables performance evaluation across multiple temporal scales. During model training, both the input (backward) and output (forward) time windows were set to 6 steps, allowing the model to capture sufficient temporal context from past observations and generate multi-step forecasts.

We trained the model for 100 epochs using the complete training dataset. Each training batch contains 32 samples for parameter updates and gradient computation. The expressiveness and complexity of the model are determined by the number of hidden units in the feed-forward layers, which was set to 256. The embedding dimension was set to 64 to provide a balanced representation capacity without overfitting. The AdamW optimizer was used for training, with an initial learning rate of 0.0003. A weight decay of 1e-4 was applied to prevent overfitting and enhance generalization. To promote stable convergence in the later stages of training, we employed a cosine annealing learning rate scheduler that gradually reduces the learning rate throughout the training process. The model's architecture includes multiple Transformer encoder layers, each consisting of a multi-head self-attention mechanism and a position-wise feed-forward network. We configured the model with 3 encoder layers and set the number of attention heads to 4, ensuring that each head has the same dimensions. Dropout regularization with a rate of 0.1 was applied within the Transformer encoders to further prevent overfitting. The configuration of key parameters is summarized in Table I.

TABLE I
MODEL PARAMETER SETTING

No.	Parameter	Value	Description
1	Input time steps	6	Historical steps used for prediction
2	Output time steps	6	Forecasting horizon
3	Feed-forward dimension	256	Internal layer size in Transformer encoder
4	Embedding dimension	64	Feature representation size per time step
5	Number of attention heads	4	Parallel attention subspaces
6	Encoder layers	3	Number of encoder layers
7	Initial learning rate	0.0003	With cosine annealing scheduler
8	Batch size	32	Training samples per batch
9	Epochs	100	Total training iterations
10	Dropout rate	0.1	Applied inside Transformer encoder

C. Comparison and Analysis of Experimental Results

i. Error Analysis

We conducted a series of tests to evaluate the performance of the proposed model and compare it with other models. Three evaluation metrics, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE), were used to comprehensively evaluate the predicted performance of each model. Specifically, MAE assesses the average deviation; MSE emphasizes larger errors due to the squaring operation; and RMSE, as the square root of MSE, provides an interpretative measure in the same unit as the predicted variable, making it particularly suitable for assessing overall predicted accuracy and the presence of significant errors. All models were trained and tested on the same dataset.

MAE is the difference between the predicted value and the actual value. The formula is presented in Equation (15).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{15}$$

In this context, y_i represents the actual value, $\hat{y}_i$ indicates the predicted value, and n denotes the sample size.

RMSE represents the square root of the squared error between the predicted value and the actual value, placing greater emphasis on the impact of larger errors. The calculation formula is shown in Equation (16). The smaller the RMSE value, the closer the model's predicted results are to the actual values, and the higher the model's predicted accuracy.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

MSE uses the square of the difference between the predicted and actual values to determine how effectively a model can predict. The squared error penalizes larger predicted errors, giving them greater weight in the overall error, so MSE is more sensitive to outliers. The smaller the MSE value, the closer the model's predictions are to the actual value. The formula is presented in Equation (17).

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{17}$$

To verify the accuracy of the model predictions, we compare the newly constructed DTW-Attention model with the classic deep learning geomagnetic storm predicted model. The models as mentioned above are detailed as follows:

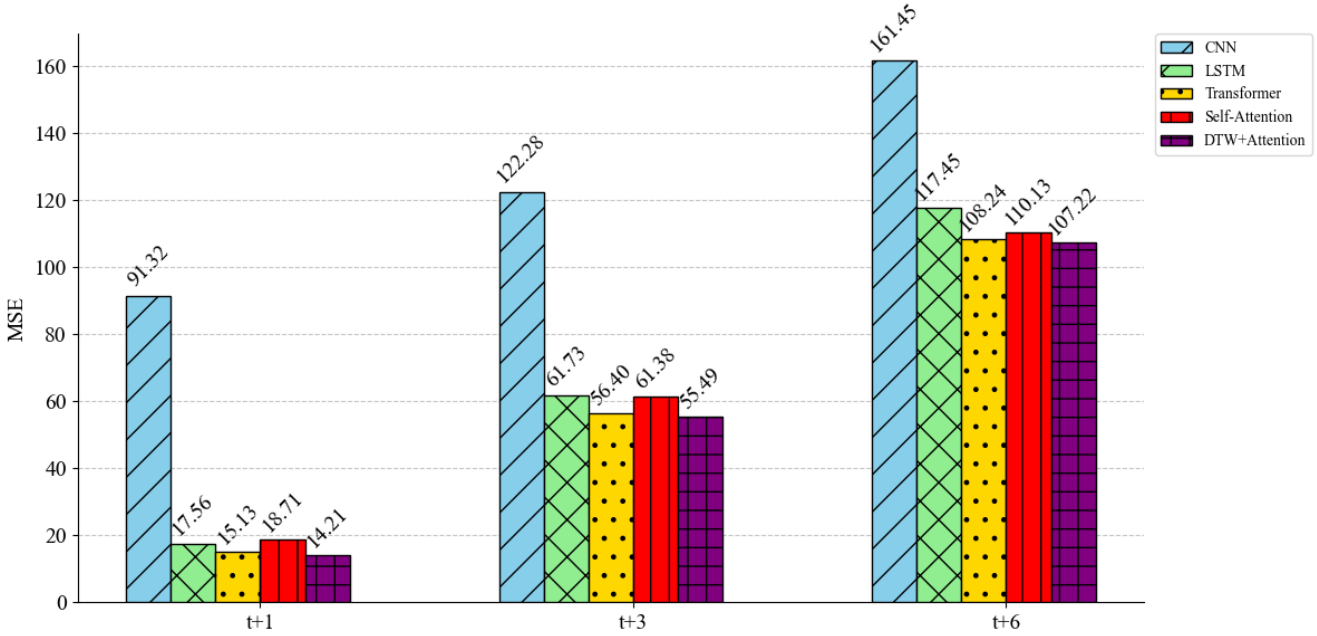
1) CNN: The CNN model has strong feature extraction capabilities and is particularly suitable for processing time series data with local dependencies. In this study, we used a three-layer one-dimensional convolutional structure combined with residual connections, batch normalization, and dropout mechanisms to effectively enhance the robustness of the model and reduce the risk of overfitting.

2) LSTM: This study introduces LSTM network, a type of RNN, as a comparison model. LSTM can effectively capture long-term dependencies in time series through its gating mechanism, making it particularly suitable for geomagnetic disturbance data with time lag and dynamic non-linear characteristics.

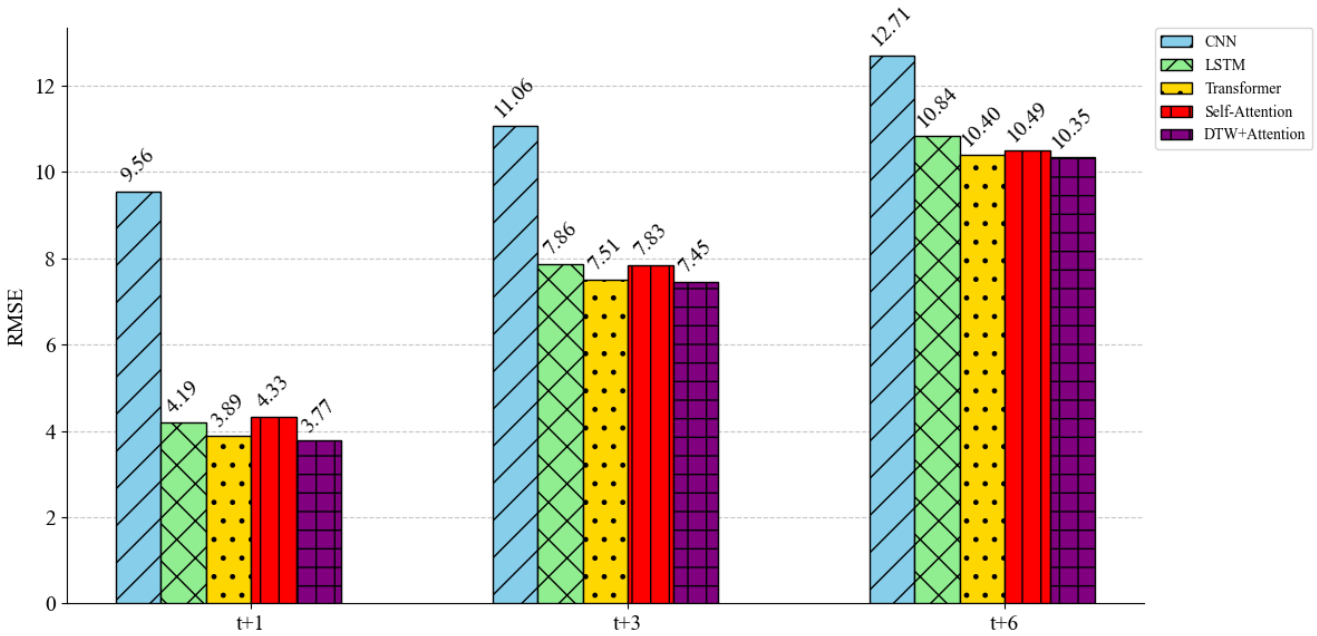
3) Transformer: Transformer is originally used for natural language processing tasks, and Self-Attention mechanism has obvious advantages in modelling long-distance dependencies and capturing global features. The purpose of introducing the model is to evaluate whether it has advantages over traditional RNN when processing multivariate spatial geomagnetic data, especially in terms of long predicted steps, such as t+6, where it can demonstrate stronger generalization and stability.

4) Self-Attention: The model automatically identifies historical moments in the input sequence that contribute most to future predictions by introducing learnable temporal attention weights, and integrates the information using weights. The core advantage of the Attention model lies in its focus extraction mechanism. It performs weighted summation on the input sequence through a set of learnable weights. This model can effectively highlight the most critical parts of historical observations.

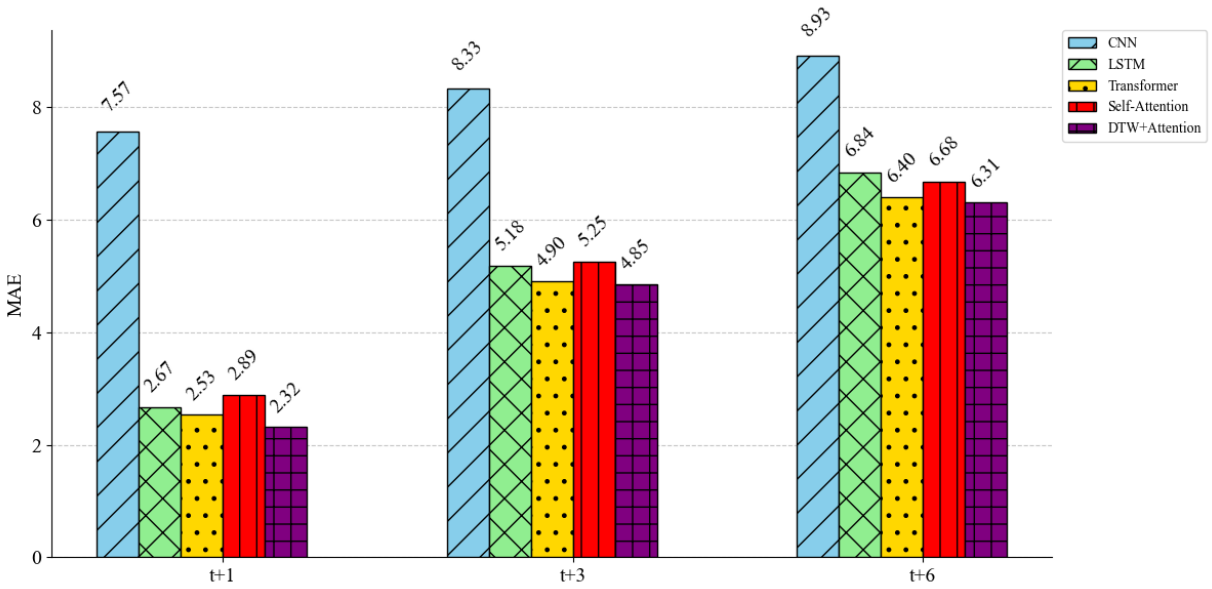
Next, we compare and analyze CNN, LSTM, Transformer, Self-Attention, and the DTW-Attention model we constructed based on three performance metrics: MSE, RMSE, and MAE.



(a) Comparison of MSE for different models at different time steps



(b) Comparison of RMSE for different models at different time steps



(c) Comparison of MAE for different models at different time steps

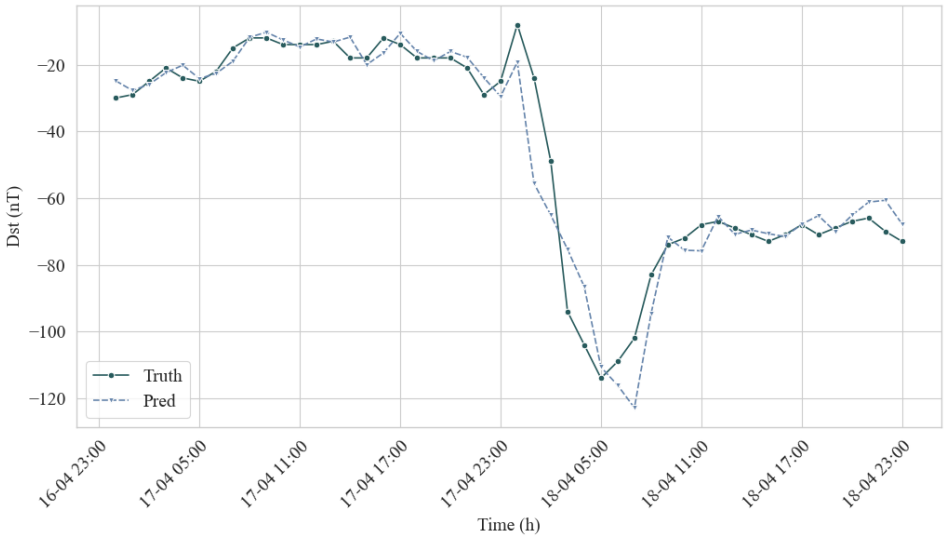
Fig.4. Comparison of errors among different models

As shown in Fig. 4(a), in the short-term prediction at t+1, all models except CNN achieved low MSE values. In particular, the DTW-Attention model had the lowest MSE, indicating its higher accuracy in capturing early changes in geomagnetic storm signals. This can be attributed to the DTW module's capability to align critical time points and the attention mechanism's capacity to emphasize salient features. In contrast, the CNN model had the highest MSE, indicating that it has significant limitations in processing strongly time-dependent data and is unable to accurately capture the early trends of geomagnetic storms. The predicted results at the t+3 time steps indicate that the LSTM and Transformer models perform stably, with their MSE significantly better than that of the CNN and slightly better than that of the Attention model. However, the DTW-Attention model still maintains the lowest error, indicating that it still has strong fitting ability for the development trend of medium-term geomagnetic storms. As the predicted time step deepens, the overall MSE values of each model increase, and the difficulty of long-term predicted significantly increases. At this point, the CNN error increases significantly, and it is almost impossible to provide effective prediction results. DTW-Attention still maintains the lowest MSE at t+6, indicating that it has a clear advantage in handling long-term, deformed time series dependencies. The model uses DTW to align the structure of key change points, effectively reducing cumulative errors in long-term predictions. Combined with the attention mechanism, it further enhances the model's ability to perceive key time features. From the data in Fig. 4(b) and 4(c), it can be analyzed that Transformer and Self-Attention models rely on attention mechanisms to dynamically allocate the importance of temporal information, and are more advantageous than traditional LSTM in capturing sudden changes in geomagnetic storm signals. Although CNN has a simple structure, it lacks a time modelling mechanism and therefore does not have an advantage in this task. The DTW-Attention model combines the time alignment of DTW with the global modelling capabilities of the Self-Attention mechanism, significantly reducing short-term and long-term errors. It is particularly suitable for processing signals such as geomagnetic storms, which exhibit both temporal deformation and sudden changes.

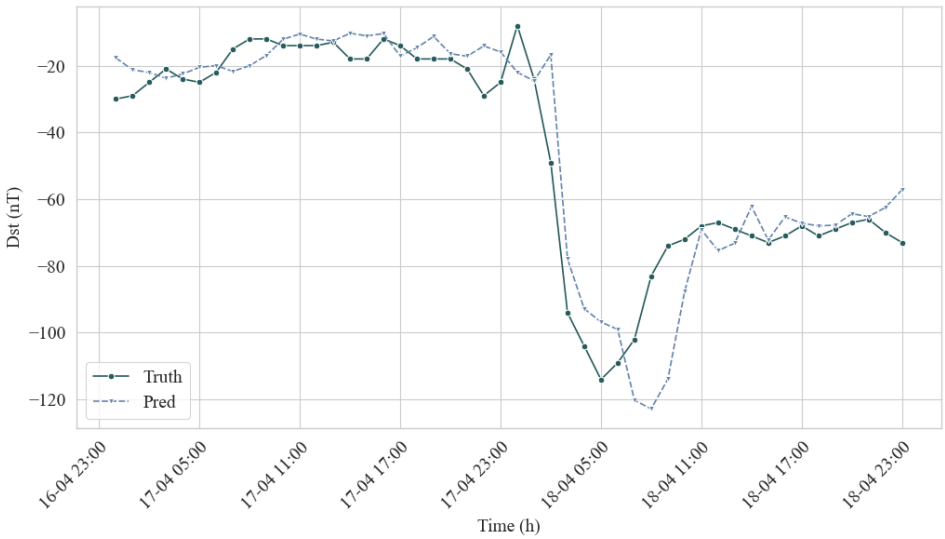
ii. Performance Evaluation

To evaluate the tracking capability of this model during geomagnetic storms, we utilized multi-line time series plots to illustrate the Dst values of both the observed (Truth) and predicted (Pred) data. Fig. 5's horizontal axis indicates time, while the vertical axis shows the Dst index. The observed Dst values, labeled as "Truth," are depicted as a solid dark green line with circular markers, clearly illustrating the actual progression of the geomagnetic storm. In contrast, the predicted Dst values, labeled as "Pred," are represented by a blue dashed line with triangular markers.

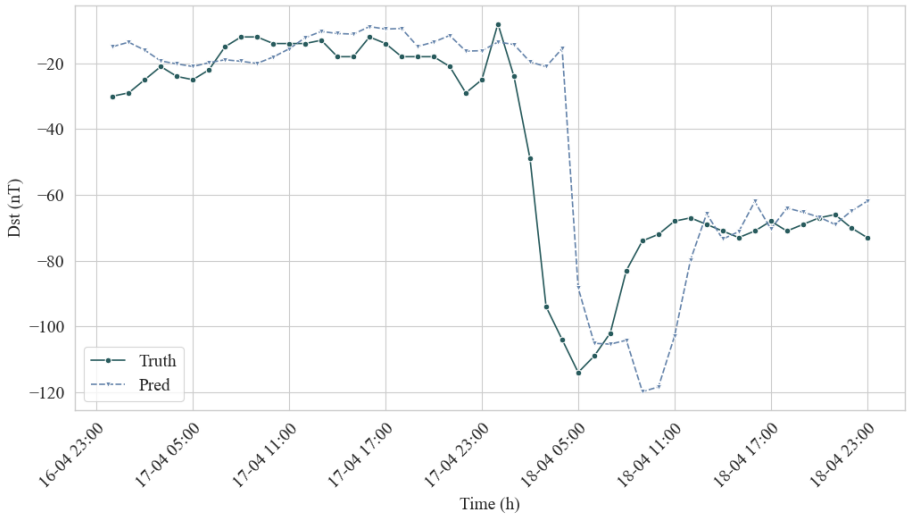
Fig. 5 shows the predicted results of the Dst index versus the observed value for the forecast time spans of t+1h hour, t+3h hour, and t+5h hour from 17 to 19 April 2001, respectively. Figure 5(a) displays the short-term predictions at t+1h hour. These have a high degree of fitting in terms of how well the estimates match the observed values, especially the Dst index's smooth change from 5:00 on the 17th to 23:00 on the 17th. Although there was some divergence and a sharp fall in Dst index between 23:00 on the 17th and 6:00 on the 18th, the general trend was still discernible. In the Fig. 5(b) t+3h hour forecast display, the forecast deviation from the observed value increases compared to the t+1h hour, especially in the phase where Dst index falls sharply; the model's ability to capture the trend is weakened. It indicates the observational uncertainty of the medium-term forecast increases. The issue of forecast delay is illustrated by the rapid changes in Dst index and the significant model forecast bias depicted in Fig. 5(c) t+5h hour forecasts, primarily occurring from 5 a.m. to 11 a.m. on the 18th. We may disregard the last six sub-figures, as they solely provide error comparisons for specific time steps and date ranges.



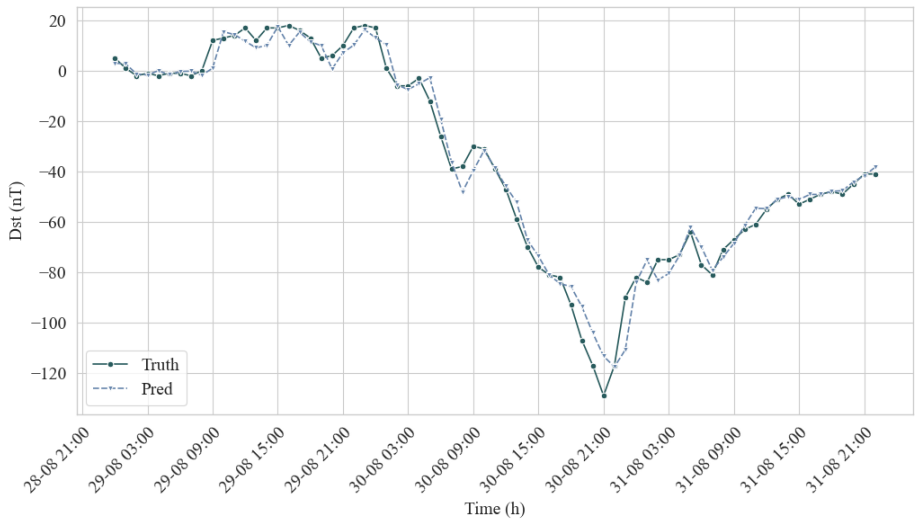
(a) Forecasting horizon t+1h (2001-04-17~04-19)



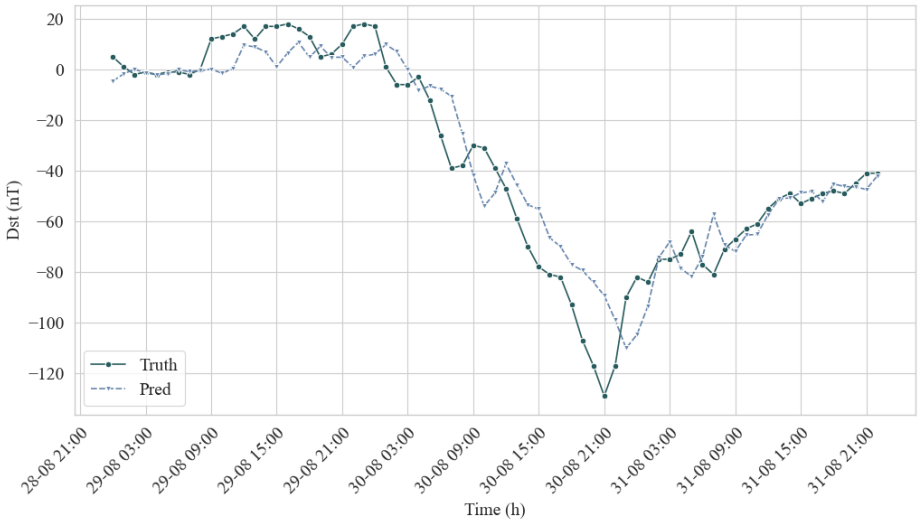
(b) Forecasting horizon t+3h (2001-04-17~04-19)



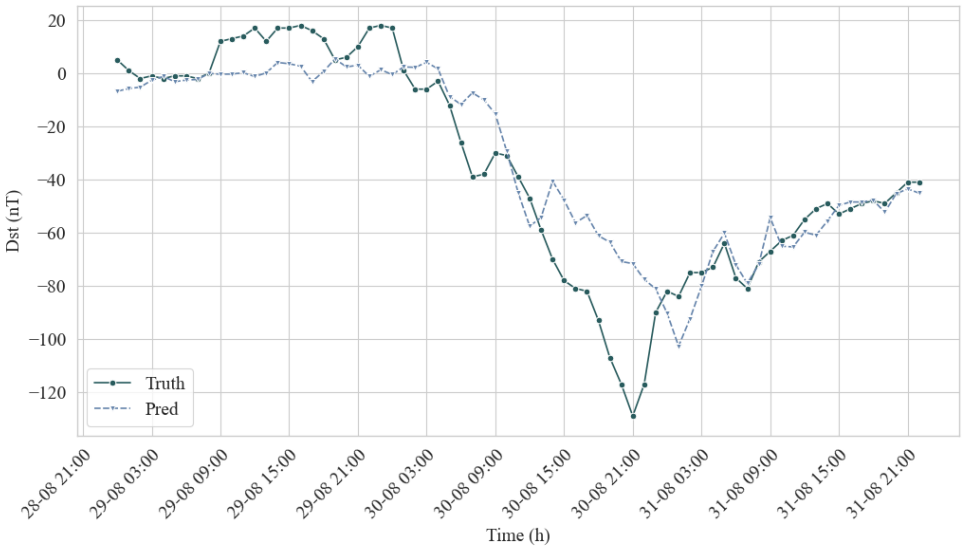
(c) Forecasting horizon t+5h (2001-04-17~04-19)



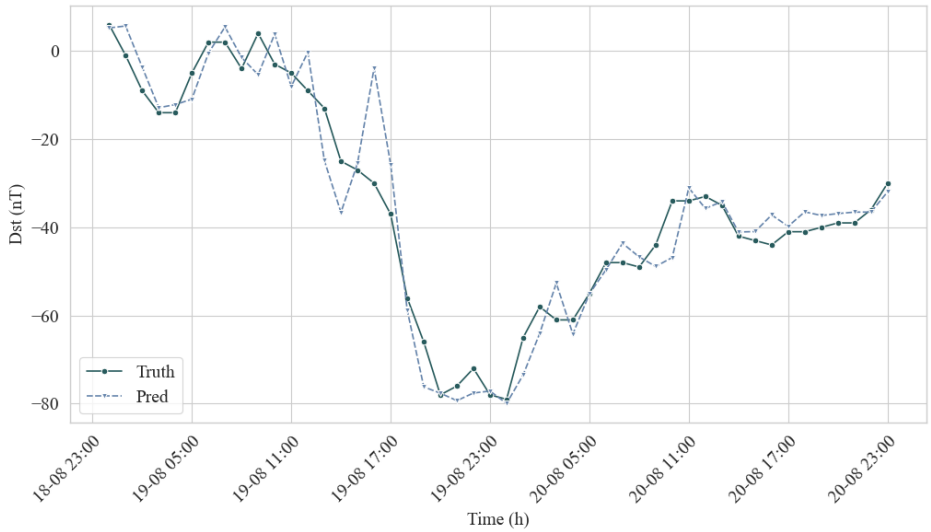
(d) Forecasting horizon t+1h (2004-08-29~09-02)



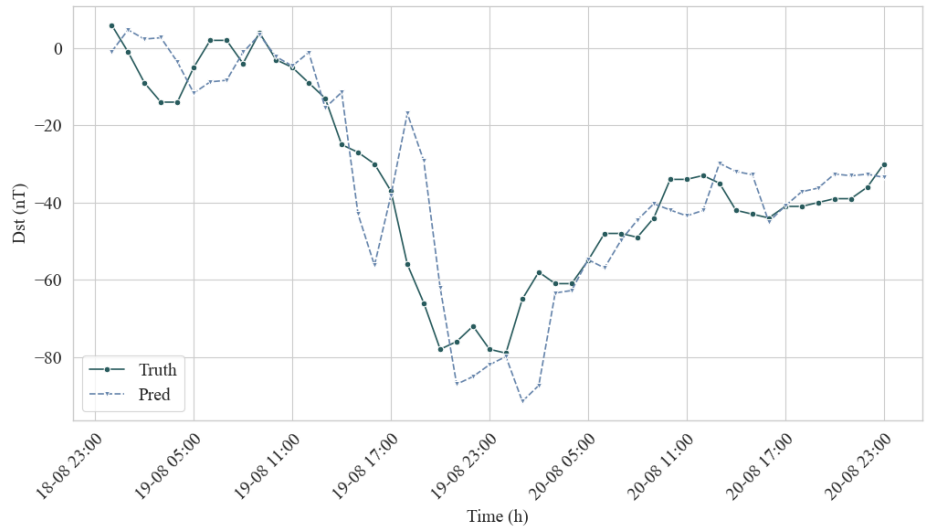
(e) Forecasting horizon t+3h (2004-08-29~09-02)



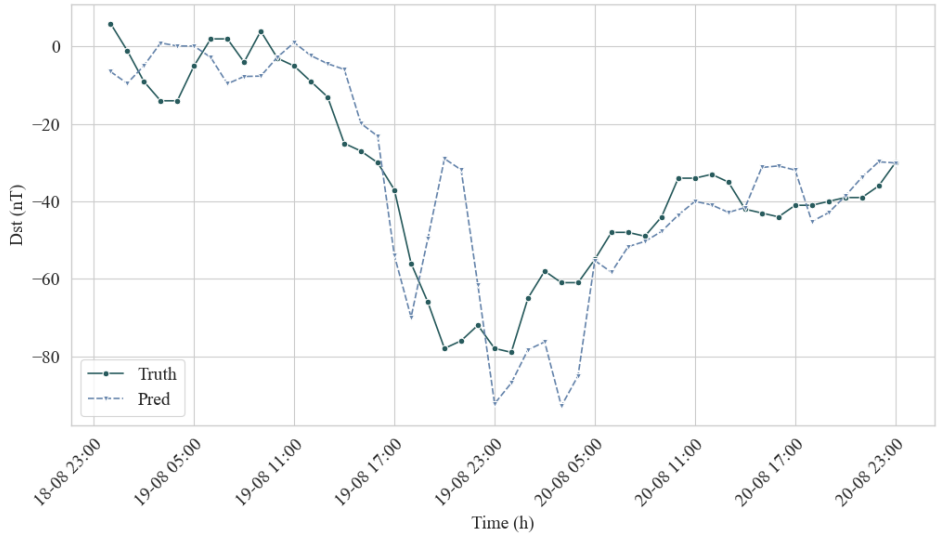
(f) Forecasting horizon t+5h (2004-08-29~09-02)



(g) Forecasting horizon t+1h (2006-08-19~08-21)



(h) Forecasting horizon t+3h (2006-08-19~08-21)



(i) Forecasting horizon t+5h (2006-08-19~08-21)

Fig.5. The predicted results of the Dst index versus the actual value for the forecast time steps

Overall, the figure allows us to visualize the range of differences between model predictions and actual geomagnetic storm intensities, as well as the times when storms are strongest. The data results indicate that the DTW-Attention model's predictions, exhibits smaller errors compared to the observed values, particularly when the predicted lengths are 1-hour and 3-hour. The model's ability to generalize across various geomagnetic storm events is evidenced by its superior predicted accuracy, which is significantly superior to other models. The visualization results also indicate that as the forecasting time step increases, the overall error level gradually rises. However, although the error of the new model also increases gradually for long-term predictions, it remains generally smaller than that of the only DTW alignment method. The result suggests that the DTW-Attention model can effectively capture long-term dependencies, demonstrating better stability and accuracy in tasks requiring future predictions. This is because the Transformer encoder captures both short-term and long-term dependencies through a multi-attention mechanism; while the DTW-Attention distance metric aligns the time series in the embedding space, thus reducing the noise effect. The model consistently outperforms the DTW alignment method across all time steps, demonstrating strong generalization ability during various geomagnetic storm phases.

D. Trend Fitting and Bias Dynamics Analysis

During the geomagnetic storm events of April 17-19, 2001 (Fig. 5(a)-5(c)), the DTW-Attention model demonstrates accurate tracking of Dst index evolution in its t+1h forecasts. Particularly noteworthy was its performance during two critical phases: the gradual decrease from 05:00 to 23:00 UTC on April 17 and the sudden Dst depression on April 18. Both predicted and observed values exhibited highly consistent variation rates, as evidenced by their parallel slope characteristics. Predicted errors primarily occurred during the fluctuation period commencing in the early hours of April 18. Nevertheless, comparative analysis revealed the model's overall forecasting trajectory maintained demonstrably superior accuracy relative to conventional approaches. The t+1h predicted capability proved especially effective in capturing short-term variations, with optimal performance observed during periods of weak geomagnetic disturbances ($|Dst| < 30$ nT).

At the t+3h and t+5h predicted steps, the model exhibits a noticeable lag, especially during periods of rapid Dst variation, such as near 06:00 on April 18. The model's responsiveness decreases at these points, likely due to residual accumulation from long-range dependency modeling. Nevertheless, even at t+5h, the DTW-Attention model effectively captures the overall trend. This is particularly evident during the recovery phase on April 19, when the predicted values rebound at a rate similar to the actual Dst values. Although medium-term predictions show some fluctuations, they still reflect the overall pattern and are suitable for early warning purposes.

In the August 2004 and 2006 events (Fig. 5(d)-5(f) and 5(g)-5(i)), the DTW-Attention model accurately identifies the inflection points of the Dst index during both the decline and recovery phases. However, the deviation from peak values increases with longer predicted horizons. For instance, in the t+5h forecast of the 2006 event, the predicted nadir is delayed. This indicates a response lag during phases of strong perturbations and rapid Dst decline. These findings suggest that integrating a local fitting sub-network or a multi-scale mechanism could enhance the model's capacity to handle high-frequency perturbations in future studies. In long-term forecasts, the model demonstrates the moderate ability to track overall trends. However, the presence of larger errors and delayed responses highlights the need to enhance the model's long-term memory and nonlinear modeling capabilities.

Observational uncertainties in geomagnetic events directly affect predicted confidence. During peak geomagnetic disturbances, elevated observational errors pose dual challenges for model training and forecasting. Nevertheless, these conditions highlight the model's robustness in handling highly noisy input data. The predicted system keeps errors within acceptable thresholds even under extreme conditions.

E. Theoretical Support and Comparative Analysis

The proposed model integrates the temporal alignment capabilities of DTW with the global sequence modeling power of the Self-Attention mechanism. This integration is theoretically supported by the complementary characteristics of the two components.

DTW is widely used to measure the similarity between temporal sequences that may vary in speed or phase. However, classical DTW operates directly on raw signal values, which makes it vulnerable to noise, scale variations, and phase distortion. Moreover, it lacks flexibility in handling long-range dependencies and often results in sub-optimal alignments due to the greedy nature of its path search. In contrast, embedding the sequences into a semantic space via a Transformer encoder allows the DTW operation to function on high-level representations. These representations are more robust to local fluctuations and capable of capturing the underlying structure of geomagnetic storm evolution. Therefore, the alignment performed in the learned feature space alleviates the limitations of traditional DTW, especially under non-stationary conditions. On the other hand, while Self-Attention excels in modeling long-range dependencies, it lacks temporal alignment awareness. Its attention matrix is symmetric and global by nature, which may lead to over-smoothing or dilution of critical transitions in the sequence. By coupling DTW alignment with attention mechanisms, the proposed model incorporates inductive bias for temporal sequencing, enabling more precise matching of rapid changes and storm peaks. Additionally, the attention weights act as soft constraints on DTW paths, effectively smoothing path fluctuations and avoiding local alignment traps.

From a complexity perspective, although the introduction of DTW increases the computational cost to $O(n^2)$, the model benefits from the parallelization of Transformer encoders and the sparsity of learned attention matrices. Recent long-sequence forecasting models such as Informer [22], Autoformer [23], and FEDformer [24]

propose alternative modeling strategies. Informer adopts ProbSparse Attention to reduce computational cost by selecting top-*k* queries, while Autoformer introduces trend-seasonal decomposition modules, and FEDformer enhances temporal modeling through frequency-domain filtering. These models are particularly effective for tasks with stable periodic structures and stationary signals.

In contrast, the DTW-Attention model focuses on alignment-aware modeling in the time domain, which is better suited to non-periodic, abrupt variations often observed in geomagnetic storm sequences. Rather than assuming decomposability or frequency regularity, our model incorporates explicit alignment mechanisms via DTW and flexible dependency capture through Self-Attention. This enables more robust learning under temporal deformation and phase shifts.

While these state-of-the-art (SOTA) models represent powerful alternatives for long-term sequence forecasting, their performance in irregular geomagnetic storm prediction remains to be fully evaluated. Future work will involve extending our experiments to include these models for a more comprehensive empirical comparison.

V. Conclusion

This paper presents a geomagnetic storm predicted model based on the DTW-Attention mechanism, which is compared with the traditional DTW alignment method in processing geomagnetic storm time-series data. The experimental results indicate that the DTW-Attention model outperforms the only DTW alignment method across several evaluation metrics, particularly showing significant advantages in capturing long-term dependencies and complex temporal data. The DTW-Attention model captures long-range dependencies by assigning time-varying attention weights. As a result, predicted accuracy and stability improve, especially for long-term forecasts. Visual analysis underscores the DTW-Attention model's effectiveness in capturing both short-term trends and long-term dynamics in geomagnetic storm sequences. Based on a comparison between the predicted results and actual observation data, the model may be able to improve the accuracy of short-term predictions while maintaining the strong robustness of long-term predictions.

While the DTW-Attention model has demonstrated strong performance in geomagnetic storm prediction, there is still room for improvement. Future studies can attempt more complex variants of the mechanism, such as multi-scale Self-Attention mechanisms and LSTM, to enhance the model's predicted ability over a longer period. Additionally, combining techniques like convolutional neural networks or graph neural networks may further improve the model's feature extraction and spatial-temporal modeling capabilities. Future research may further explore online learning mechanisms to enable real-time geomagnetic storm forecasting for practical applications.

References

[1] Adrian Tasistro-Hart, Alexander Grayver, and Alexey Kuvshinov, et al., "Probabilistic Geomagnetic Storm Forecasting via Deep Learning," Journal of Geophysical Research: Space Physics, vol. 126, https://doi.org/10.1029/2020JA028228, 2021.

[2] Iris Yan, "Early Prediction of Geomagnetic Storms by Machine Learning Algorithms," https://doi.org/10.48550/arXiv.2401.10290, 2024.

[3] X. Huang Z. R. Zhao, and Y. F. Zhong, et al., "Short-term solar eruptive activity prediction models based on machine learning approaches: A review," Science China Earth Sciences, vol .67, no. 12, pp 3727-3764, 2024.

[4] Stumpo M, Benella S, and Laurenza M, "Open issues in statistical forecasting of solar proton events: A machine learning perspective," Space Weather, vol. 19, no. 10, https://doi.org/10.1029/2021SW002794, 2021.

[5] Krizhevsky A, Sutskever I, and Hinton G E, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp 84-90, 2017.

[6] Siciliano, F., Consolini, G., and Tozzi, R., "Forecasting SYM-H Index: A Comparison Between Long Short-Term Memory and Convolutional Neural Networks," Space Weather, vol. 19, no. 2, https://doi.org/10.1029/2020SW002589, 2021.

[7] E. Camporeale, "The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting," Space Weather, vol. 17, no. 8, pp 1166-1207, 2019.

[8] Carlos Escobar Ibáñez, Daniel Eduardo, and Florencia Luciana Castillo, et al., "Forecasting Geomagnetic Storm Disturbances and Their Uncertainties Using Deep Learning," Authorea, https://doi.org/10.22541/essoar.167839956.63152489/v1, 2023.

[9] Engell A J, Falconer D A, and Schuh Mi, "SPRINTS: A framework for solar-driven event forecasting and research," Space Weather, vol. 15, no. 10, pp 1321-1346, 2017.

[10] Mishka Alditya Priatna, and Esmeralda C., "Precipitation prediction using recurrent neural networks and long short-term memory," Telecommunication, Computing, Electronics and Control, vol. 18, No. 5, pp 2525-2532, 2020.

[11] M. Cristoforetti, R. Battiston, and A.Gobbi, et al., "Prominence of the Training Data Preparation in Geomagnetic Storm Prediction Using Deep Neural Networks," Scientific Reports, vol. 12, no. 1, https://doi.org/10.1038/s41598-022-11721-8, 2022.

[12] Hakan Uyanık, Erman Şentürk, and Muhammed Halil Akpınar, et al., "A Multi-Input Convolutional Neural Networks Model for Earthquake Precursor Detection Based on Ionospheric Total Electron Content," Remote Sensing, vol. 15, no. 24, https://doi.org/10.1029/2022SW003231, 2023.

[13] A. Yasser, W. Jason, and B. Prianka, et al., "Forecasting the Disturbance Storm Time Index with Bayesian Deep Learning," Proceedings of the International Florida Artificial Intelligence Research Society Conference, vol. 35, https://doi.org/10.32473/flairs.v35i.130564, 2022.

[14] Yasser Abduallah, Jason T. L. Wang, and Prianka Bose, et al., "A Deep Learning Approach to Dst Index Prediction," https://doi.org/10.48550/arXiv.2205.02447, 2022.

[15] Zhihong Lv, Rui Sun, and Xin Liu, "Evaluating the effectiveness of Self-Attention mechanism in tuberculosis time-series forecasting," BMC Infectious Diseases, vol. 24, no. 1, pp 1377-1389, 2024.

[16] Junhong Chen, Hong Dai, Shuang Wang, and Chengrui Liu, "Improving Accuracy and Efficiency in Time Series Forecasting with an Optimized Transformer Model," Engineering Letters, vol. 32, no. 1, pp 1-11, 2024.

[17] S Wang, H Dai, and L Bai, et al., "Temporal Branching-Graph Neural ODE without Prior Structure for Traffic Flow Forecasting," Engineering Letters, vol. 31, no. 4, pp 1534-1545, 2023.

[18] NASA. (2025). OminiWeb:[GODDARD SPACE FIGHT CENTER Space Physics Data Facility]. https://omniweb.gsfc.nasa.gov/ow.html

[19] J. H. King, "Solar Wind Spatial Scales in and Comparisons of Hourly Wind and ACE Plasma and Magnetic Field Data," Journal of Geophysical Research, vol. 110, no. A2, https://doi.org/10.1029/2004ja010649, 2005.

[20] Yuxiang Peng, Jianyong Lü, and Saiju GU, "Application of Support Vector Machine to the Forecasting of Dst Index during Geomagnetic Storm," Chinese Journal of Space Science, vol. 36, no. 6, https://doi.org/10.11728/cjss2016.06.866, 2016.

[21] X D Ren, P X Yang, and D K Mei, et al., "Global Ionospheric TEC Forecasting for Geomagnetic Storm Time Using a Deep Learning-Based Multi-Model Ensemble Method," Space Weather the International Journal of Research and Applications, vol. 21, no. 3, pp 3231-3248, 2023.

[22] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, et al., “Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp 11106–11115, 2021.

[23] XingRui Fan, and Yuancheng Li , “Short-term power load forecasting based on improved Autoformer model,” Dianli

Zidonghua Shebei/Electric Power Automation Equipment, vol. 44, no. 4, pp 171-177, 2024.

[24] Lei Ge, Qiwei Huang, and FengShuang Zhu, et al., "Advanced time series forecasting for commodities: Insights from the FEDformer model," Energy Economics, vol. 147, https://doi.org/10.1016/j.eneco.2025.108513, 2025.