Space Weather

RESEARCH ARTICLE
10.1029/2024SW004121

Channel Mixer Layer: Multimodal Fusion Toward Machine Reasoning for Spatiotemporal Predictive Learning of Ionospheric Total Electron Content

Peng Liu[1], **Tatsuhiro Yokoyama**[1], **Takuya Sori**[1], **and Mamoru Yamamoto**[1]

[1]Research Institute for Sustainable Humanosphere, Kyoto University, Uji City, Japan

Key Points:

- Channel mixer layer is proposed to improve the Total Electron Content (TEC) prediction accuracy of existing models by multimodal fusion of external dependence factor
- The largest standard TEC data set and comprehensive software for spatiotemporal predictive learning is proposed to ensure the fair comparison
- Experiment results show that the proposed method has the highest TEC prediction accuracy, real-time speed and best machine reasoning ability

Correspondence to:
P. Liu,
liu.peng.35a@st.kyoto-u.ac.jp.

Citation:
Liu, P., Yokoyama, T., Sori, T., & Yamamoto, M. (2024). Channel mixer layer: Multimodal fusion toward machine reasoning for spatiotemporal predictive learning of ionospheric total electron content. *Space Weather*, *22*, e2024SW004121. https://doi.org/10.1029/2024SW004121

Received 12 AUG 2024
Accepted 25 NOV 2024

Author Contributions:
Conceptualization: Peng Liu
Data curation: Peng Liu
Formal analysis: Peng Liu
Funding acquisition: Tatsuhiro Yokoyama
Investigation: Takuya Sori
Methodology: Peng Liu
Project administration: Mamoru Yamamoto
Resources: Peng Liu
Software: Peng Liu
Supervision: Tatsuhiro Yokoyama
Validation: Peng Liu
Visualization: Peng Liu
Writing – original draft: Peng Liu
Writing – review & editing: Tatsuhiro Yokoyama

Abstract

The spatiotemporal distribution of Total Electron Content (TEC) in ionosphere determines the refractive index of electromagnetic wave leading to the radio signal scintillation and deterioration. Thanks to the development of machine learning for video prediction, spatiotemporal predictive models are applied on the future TEC map prediction based on the graphic features of past frames. However, output result of graphic prediction is unable to properly respond to the external factor variations such as solar or geomagnetic activity. Meanwhile, there is still neither standard data -set nor comprehensive evaluation framework for spatiotemporal predictive learning of TEC map sequences leading to the comparisons unfair and insights inconclusive. In this research, a new feature-level multimodal fusion method named as channel mixer layer for machine reasoning is proposed that can be embedded into the existing advanced spatiotemporal sequence prediction models. Meanwhile, all performance benchmarks are accomplished on the same running environment and newly proposed largest scale data set. Experiment results suggest that the multimodal fusion prediction of existing model backbones by proposed method improves the prediction accuracy up to 15% with almost the same computational complexity compared to that of graphic prediction without auxiliary factors input, having the real-time inference speed of 34 frames/second and minimum mean absolute error of 0.94/2.63 TEC unit during low/high solar activity period respectively. The channel mixer layer embedded models can respond to the variations of auxiliary external factors more correctly than previous multimodal fusion methods such as concatenation and arithmetic, which is regarded as the evidence of state-of-the-art machine reasoning ability.

Plain Language Summary

Total Electron Content (TEC) refers to the total number of free electrons along the electromagnetic wave path while penetrating ionosphere. The remote sensing of ionospheric TEC is accomplished by calculating the group delays (phase advances) of radio signal between satellites and ground receivers. Recently the prediction of global TEC maps has been the research highlights due to its great significance for the satellite communication improvement. However, the spatiotemporal sequence prediction models used in previous researches are not state-of-the-art without focusing on the modeling of multimodal fusion due to the input of different kinds of external factor data such as solar and geomagnetic activity indices. In this research, a new multimodal fusion method named as channel mixer layer for machine reasoning is proposed that can be embedded into the existing advanced spatiotemporal sequence prediction models while inputting auxiliary data. Experiment results under the same newly proposed data -set and software suggest that the proposed multimodal fusion method improves the prediction accuracy dramatically compared to the graphic prediction without auxiliary data input. Channel mixer layer shows the best machine reasoning ability by correctly responding to the variations of auxiliary factors under different external conditions.

1. Introduction

Ionosphere is the upper atmosphere from 60 to 1,000 km altitude where neutral particles are partially ionized as plasma by solar radiation. The electromagnetic wave refraction in the ionosphere caused by charged particles is one of the main reasons for the satellite signal scintillation and radio quality deterioration. The Total Electron Content (TEC) between a pair of radio transmitter and receiver is defined as the total number of free electrons (TEC Unit: 1 TECU = 10^{16} electrons/m^2) contained in a cylinder (1 m^2 cross section) linking them (H. Wang et al., 2024). The remote sensing of ionospheric TEC has been carried on since 1995 by calculating the group delays (phase advances) of radio signal between Global Positioning System (GPS) satellites and hundreds of worldwide ground receivers (J. Liu et al., 2011). However, considering some technical problems such as the incomplete data due to observation interruptions before 1998 (Z. Wang, Wu, et al., 2023), sampling time shifting

from odd to even hours in 2002 (Ren et al., 2022) and the change of sampling interval from two to 1 hour in 2014 (Luo et al., 2023), the data sets of global TEC maps are partially and differently used in previous researches. In this research, a standard data set with the largest scale is released where the global TEC maps and their corresponding external factors for diurnal, seasonal, spatial, solar and geomagnetic activity dependence are packaged together after removing the redundant descriptions in origin text data files.

The prediction models for global ionosphere TEC can be roughly divided into two categories: physical and empirical models (Shi et al., 2022). In terms of physical models, the fully coupled momentum, energy, and continuity equations for both neutral and ion particles are solved to derived the global TEC maps (Codrescu et al., 2012). However, all physical prediction models are established based on the first principles that assume the steady states for physics and chemistry ignoring the complex variations in ionosphere (Ridley et al., 2006). Thus, there is bias error that overestimates or underestimates the TEC predictions (Perlongo et al., 2018). Empirical models can also be divided into two categories: function-based and learning-based models. Function-based models such as International Reference Ionosphere (IRI, Rawer et al. (1978)) and NeQuick (Nava et al., 2008) always fit the spatiotemporal distribution of ionospheric parameters with the complex functions. The predictions of these models are significantly lower than observations due to the exclusion of plasmaspheric TEC contribution (Bilitza et al., 2022). Recently, learning-based models have been the research highlights due to the merits of high accuracy and real-time speed. However, the spatiotemporal sequence prediction models used in previous researches are not state-of-the-art (Ruwali et al., 2021; Z. Chen et al., 2022). Meanwhile, the accuracy comparison of different models in previous researches may be unfair due to the different computational consumption (Xiong et al., 2021). In this research, the TEC prediction performances of 11 advanced models in machine learning research field are evaluated under the same modular and extensible framework with the comparable model configuration.

Unlike the previous researches which propose the whole model completely (Srivani et al., 2019), this research focuses on improving the existing mature models by the multimodal fusion between predictive channel (global TEC maps) and auxiliary channel (external factors). Without the appropriate multimodal fusion methods, the model prediction accuracy may degrade with more auxiliary channels input according to the experiment results. The existing multimodal fusion methods can be roughly categorized as feature-level (or early) fusion and decision-level (or late) fusion (Baltrušaitis et al., 2019). In terms of decision-level fusion, several submodels are trained with the classified data separately under different conditions such as different time and solar activity. The prediction results of different trained submodels are selected to make the final decision according to a certain criterion including voting scheme (Nigusie et al., 2024), maximum likelihood (Li et al., 2024), bayes theorem (Karatay & Gul, 2023) and ensemble learning (such as Gradient Boosting algorithm, Zhou et al. (2024)). However, The prediction speed is slow due to the large model size and repetitive iteration. Meanwhile, the association between different single submodels is ignored so that it is unsuitable for TEC prediction where the external factors have strong correlation. In terms of feature-level fusion, the common implementations is applying arithmetic (e.g., weighted sum (L. Liu et al., 2020) and multiplication (Y. Wang, Wu, et al., 2023)) or concatenation (Xu et al., 2024) operation on the extracted features of predictive and auxiliary channels. These methods can be regarded as the multimodal fusion after the feature extraction. In this research, a new fusion-before-extraction multimodal fusion framework named as channel mixer layer for machine reasoning is proposed that can be embedded into the existing advanced spatiotemporal sequence prediction models.

2. Data Set

Since 1995, the global TEC maps have been provided by the Center for Orbit Determination in Europe (CODE) based on the signal delay between the globally distributed ground receivers and satellites (Dow et al., 2009). However, the start time of observation changed from Universal Time (UT) 1:00 to 00:00 on 3 November 2002 and the sampling interval of observation data changed from two hours to one hour on 19 October 2014, which makes previous researches cannot make full use of observation data (Ren et al., 2022). In this research, the observation TEC maps of each day since 19 October 2014 are divided into odd and even universal time groups to ensure all sequences have the same two-hour interval so that the data-set can cover the longest time span. Meanwhile, The TEC maps are resized from 73×71 to 72×72 pixel resolution with a spatial resolution of 5° and 2.5° for longitude and latitude at 450 km ionospheric mapping shell altitude (Fu et al., 2022), because existing spatiotemporal predictive models in machine learning generally requires that the sequence frames have the same width and height size. The detailed approaches include removing the right edge column because both it and left edge

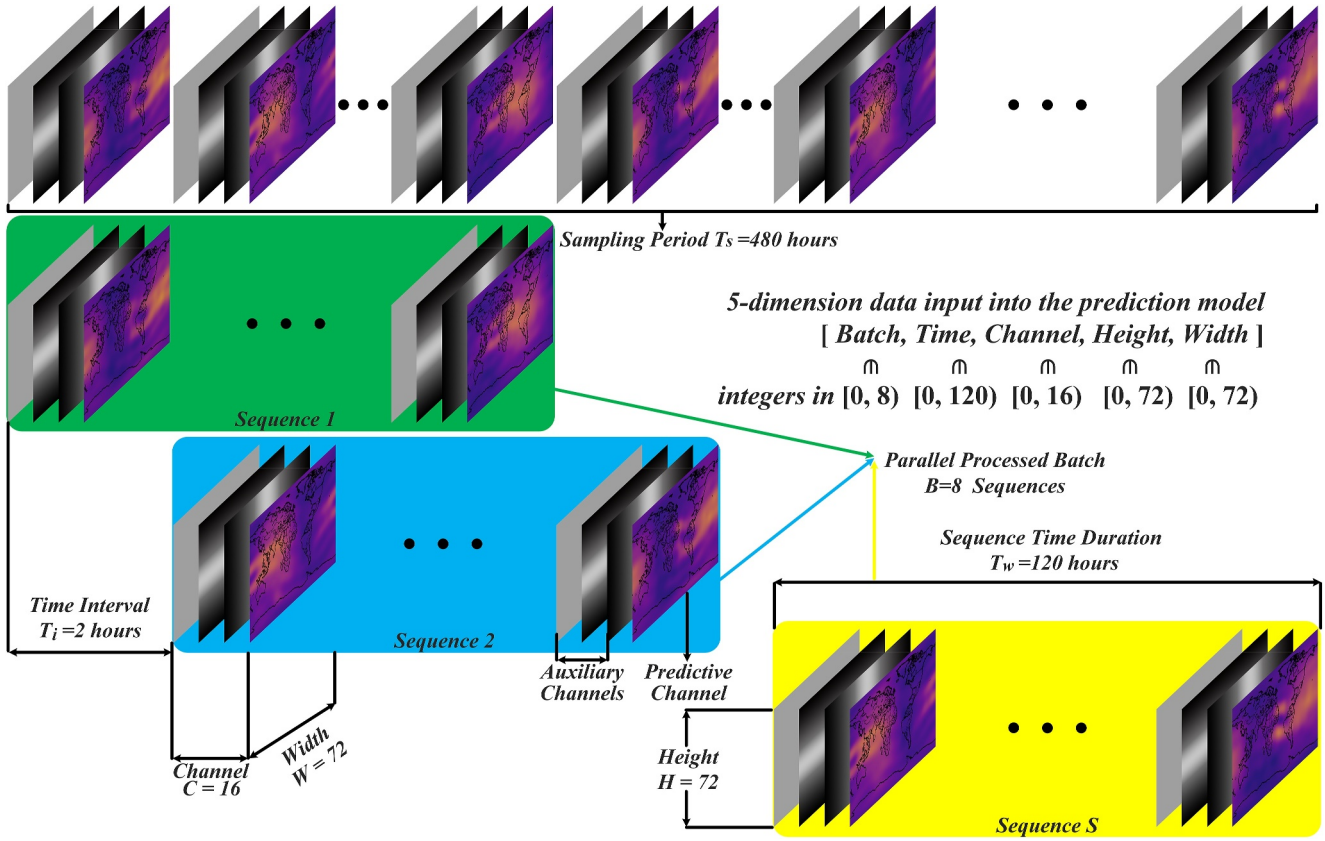

Figure 1. The sequence quantity augmentation method is used to generate the 5-dimension data for prediction models where the sequences are automatically selected by a sliding window along the observation samples.

column are the same data for the TEC at 180° longitude repeatedly, meanwhile adding a new zero-padding bottom edge row without changing the original data.

There is still no standard data set for spatiotemporal predictive learning of TEC map sequences leading to the comparisons unfair and insights inconclusive. In this research, a new standard data set with the largest scale, named as IonoElectron, is released where the global TEC maps and their corresponding external factors for diurnal, seasonal, spatial, solar and geomagnetic activity dependence provided by OMNI data set are packaged together and saved in the specified format of machine learning software for the high speed accessing. Meanwhile, to prevent the model over-fitting caused by insufficient data, the sequence quantity augmentation method as shown in Figure 1 is applied where the sequences to be input are automatically selected by a sliding window along the observation samples. Supposing the total number of the ionospheric TEC maps is $N = 141,360$ frames from 1997 to 2022 and the time interval (time resolution) between two contiguous frames is $T_i = 2$ hours, the sampling period $T_s = 480$ hours refers to the longest sequence duration that can be sampled due to the observation interruption, variable $T_w = 120$ hours $(0 < T_w \le T_s)$ refers to the time length of sliding window. Then augmented sequence quantity $S = 106,609$ can be calculated by Formula 1, which is much more than that $S = (N \cdot T_i)/T_w = 2356$ without applying sequence quantity augmentation method and is the largest scale compared to the similar researches until now.

$$S = \frac{N}{T_s}(T_s - T_w + T_i) \tag{1}$$

With the aforementioned data processing approach, each sequence for the spatiotemporal sequence prediction can be regarded as a 5-dimension tensor of [batch, time, channel, height, width], where the batch dimension refers to the sequence quantity that parallel processed by the model and the channel dimension is composed of predictive

and auxiliary channels. Each individual data in the sequence can be located by the 5-dimension coordinate system within the range as shown in Figure 1.

The sequences obtained by sequence quantity augmentation are separated into training, validation and test data sets according to the different observation time with a number portion ratio of 3:1:1. The cross validation strategy is adopted for training and validation data sets. The test data set is further grouped as the low (annually averaged $F_{10.7} < 80$ sfu) or high (annually averaged $F_{10.7} > 130$ sfu) solar activity periods. Due to the positive correlation between solar and geomagnetic activities (P. Liu et al., 2022), this categorization can also reflect the performance of the model under different geomagnetic activity conditions.

3. Methodology

3.1. Network Architecture

This research aims to enhance the machine reasoning ability of models for the relation between predictive and auxiliary channels through the multimodal fusion. To this end, we outline the development background of technical features of existing sequence prediction models, which can be roughly categorized as Recurrent Neural Networks (RNN, Elman (1990)), Convolutional Neural Networks (CNN, LeCun et al. (1989)) and Transformer (Vaswani et al., 2017). Two branches of RNN models, Long-Short Term Memory (LSTM, Hochreiter and Schmidhuber (1997)) and Gated Recurrent Unit (GRU, Chung et al. (2014)) were proposed to solve the gradient vanishing/exploding problems in simple RNN model. These models were mainly applied in temporal sequence prediction tasks such as natural language processing in the early stage. By switching from fully connected operations of LSTM model to convolutional operations to preserve spatial features, Convolutional LSTM (ConvLSTM) established spatiotemporal sequence prediction research field (SHI et al., 2015). Another seminal research, Predictive RNN (PredRNN), used different modules to learn spatial and temporal features separately (Y. Wang et al., 2017). Based on aforementioned principles, further improvements have been made by the subsequent researches, such as PredRNN++ (Y. Wang et al., 2018), PredRNNv2 (Y. Wang, Wu, et al., 2023), Memory in Memory (MIM, Y. Wang, Zhang, et al. (2019)) and Eidetic 3-dimensional LSTM (E3D-LSTM, Y. Wang, Jiang, et al. (2019)). Besides the Recurrent-based models, the recurrent-free model based on CNN is also proposed such as Simple Video Prediction (SimVP, Gao et al. (2022)). Considering the success of the attention module, recent researches such as Temporal Attention Unit (TAU, Tan et al. (2023)) and Motion Aware Unit (MAU, Chang et al. (2021)) combine the merits of Transformer and other models. These models are used as the prediction network backbone for the multimodal fusion framework proposed in this research. The detailed properties of the most diverse types of data channels and model backbones used in this research (Tang et al., 2024) are shown in Table 1.

Figure 2 shows the prediction process of the different kinds of neural networks for the predictive channel (ionospheric TEC map) with/without inputting the auxiliary channels (external factors). The aforementioned models for spatiotemporal sequence prediction require that the channel number of the input and output data should be equivalent by default because the prediction based on graphic features does not need the extra auxiliary input. In this research, we extend the functionality of input layers in the existing models to make them support any number of channels. When the machine reasoning from the auxiliary to predictive channels is enabled, the data of the predictive and auxiliary channels is fused by the multimodal fusion method before inputting into prediction network backbone. In terms of the recurrent-based models like RNN, the data of auxiliary channels at the current time step needs to be copied and put together with the prediction result at the last time step to ensure the consistency of the input. Whereas such an operation is not needed for the recurrent-free models like CNN or Transformer where all of the prediction results are output for once.

The mathematical expression of spatiotemporal sequence predictive learning can be summarized by Formula 2, where $X_{\text{in}} = \{x_{t-T_{\text{in}}+1}, \ldots, x_t\}$ and $Y_{\text{out}} = \{y_{t+1}, \ldots, y_{t+T_{\text{out}}}\}$ refer to the input and output sequences consisting of the frames of x and y during time length T_{in} and T_{out} respectively. The goal of learning procedure is to find the optimized network parameters for graphic prediction $\hat{Y}_{\text{out}}^{C_{\text{pred}}} = f\left(X_{\text{in}}^{C_{\text{pred}}}\right)$ and multimodal fusion prediction $\hat{Y}_{\text{out}}^{C_{\text{pred}}} = f\left(X_{\text{in}}^{C_{\text{pred}}}, X_{\text{in}}^{C_{\text{aux}}}\right)$ minimizing the loss function result $\mathcal{L}(\text{output}, \text{truth})$, where C_{pred} and C_{aux} refer to the number of predictive and auxiliary channels.

Table 1
The Detailed Properties of the 16-Channel Predictive or Auxiliary Data and 12-Kind Temporal or Spatiotemporal Model Backbones Used in This Research

No.	Channel	Dependence	Model	Category	Structure
1	TEC map	–	Temporal		
2	Longitude	Spatial	Prediction:		
3	Latitude	Spatial	LSTM	RNN	Fully connected LSTM cell
4	Year	Seasonal	GRU	RNN	Fully connected GRU cell
5	Day	Seasonal	Transformer	Transformer	Attention
6	Month	Seasonal	Spatiotemporal		
7	Hour	Diurnal	Prediction:		
8	F10.7	Solar	ConvLSTM	RNN	Convolutional LSTM cell
9	Sunspot	Solar	E3D-LSTM	RNN	3-Dimensional ConvLSTM cell
10	Kp	Geomagnetic	MAU	RNN + Transformer	Recurrent + Attention
11	Ap	Geomagnetic	PredRNN	RNN	Spatiotemporal ConvLSTM cell
12	Dst	Geomagnetic	PredRNN++	RNN	Deeper PredRNN + Highway
13	AE	Geomagnetic	PredRNNv2	RNN	PredRNN + Decoupled memory
14	ΣB	Geomagnetic	MIM	RNN	Differential PredRNN
15	Average$\lvert B\rvert$	Geomagnetic	SimVP	CNN	Encoder + Inception + Decoder
16	PlasmaSpeed	Geomagnetic	TAU	CNN + Transformer	Encoder + Attention + Decoder

$$\min_{f} \sum_{\text{batch}} \mathcal{L}\left(\hat{Y}_{\text{out}}^{C_{\text{pred}}} = f(X_{\text{in}}), Y_{\text{out}}^{C_{\text{pred}}}\right) \tag{2}$$

3.2. Multimodal Fusion

Traditional spatiotemporal predictive learning always focuses on predicting the future frames only according to the graphic features of past frames. However, the spatiotemporal sequences of ionospheric TEC maps have strong correlation with external factors, which represents as the diurnal, seasonal, spatial, solar and geomagnetic activity dependence (P. Liu et al., 2022). The prediction accuracy is expect to be further improved with the machine reasoning, where the relations between the predictive channel (ionospheric TEC map) and the auxiliary channels (external factors) are learned by neural networks with the suitable multimodal fusion method.

Figure 3a shows the processing flowchart of the previous multimodal fusion methods. All of the input data are normalized within the range of (0,1) due to the activation functions in neural networks. The learning process of one parameter in a certain layer of the neural network is emphasized where the calculation result of convolution or attention operation between this parameter and the input data are regarded as the extracted feature then passed to the next layer. The parameter sharing is a fundamental technique in neural networks which means that the same parameter scans over different regions and different channels instead of using different parameters. Therefore, there are dramatic gaps for the extracted features and learned parameters of different channels represented as different colors leading to informatics loss and learning unstable.

Existing multimodal fusion methods such as arithmetic and concatenation are applied on the extracted features of different channels. In terms of the arithmetic methods such as weighted sum and multiplication, the result of weighted sum method may exceed the range within (0,1). Multiplication method makes the result closer to zero when more channels are input, leading to the vanishing gradient problem. To solve these problems, normalization layers are applied so that the parameter number of the neural network increases dramatically. Meanwhile, the weight of each channel W_i is difficult to be determined for the weighted sum method. In machine learning researches, multiplication method may have the different name such as action-conditioned PredRNN (Y. Wang, Wu, et al., 2023) but essentially is same. In terms of concatenation method, extracted features of different channels are put together without extra operation. All of the relations between the predictive and auxiliary channels are determined by the neural network backbone. The original codes of model backbones that only

10.1029/2024SW004121

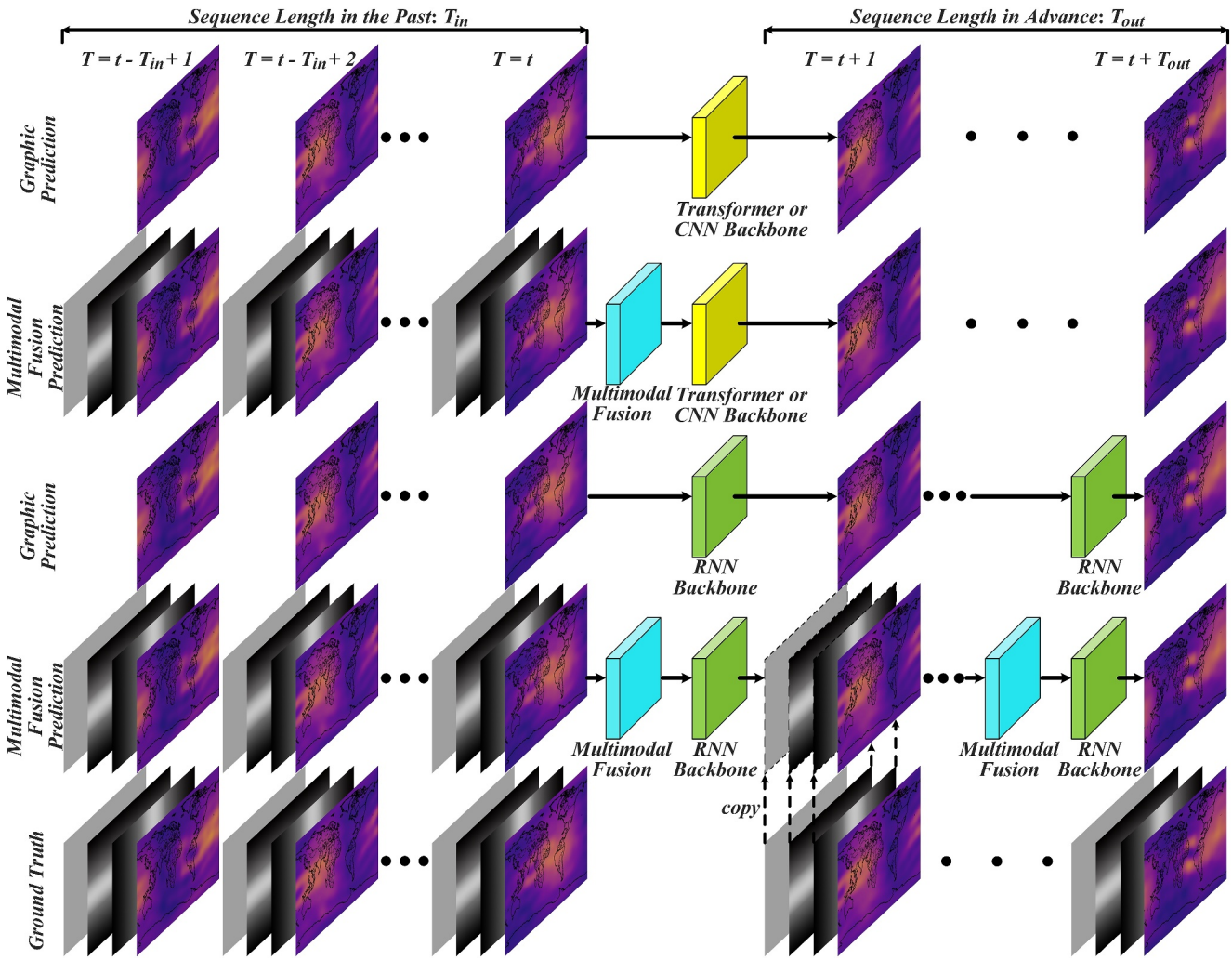

Figure 2. The flowchart of different models and different prediction categories. The difference between graphic prediction and multimodal fusion prediction is whether the auxiliary channel data is input or not. The recurrent-based models predict step by step whereas others output the prediction sequence for once.

support the graphic prediction are modified and upgraded to the multimodal fusion prediction with the concatenation method in this research.

Different from traditional methods that extracts features before multimodal fusion, the proposed mixer layer is a framework that extracts features after multimodal fusion as shown in Figure 3b. The original data of the same position of the different channels are replaced nearby with each other into the same channel so that the distribution differences between predictive and auxiliary channels can be eliminated. However, the mixer layer makes the distance between the data of predictive channel $\sqrt{C}$ times larger, where C refers to the total number of channel, leading to the decrease of the prediction accuracy. In order to maintain the same receptive field and kernel size without increasing the trainable parameters of model, a feature extraction bypass for the predictive channel without through mixer layer is set. All of the extracted features of different channels are concatenated together then input into neural network backbone for the spatiotemporal prediction.

The traditional mixer layers proposed for the object recognition (Y. Wang, Sun, et al., 2023) and temporal sequence prediction (Tolstikhin et al., 2021) contain the fully connected operations, which are superseded in the existing spatiotemporal predictive models due to the spatial feature loss. However, our mixer layer is essentially a new data transformation without the trainable parameter or fully connected operation so that it can be flexibly integrated into the existing model backbones.

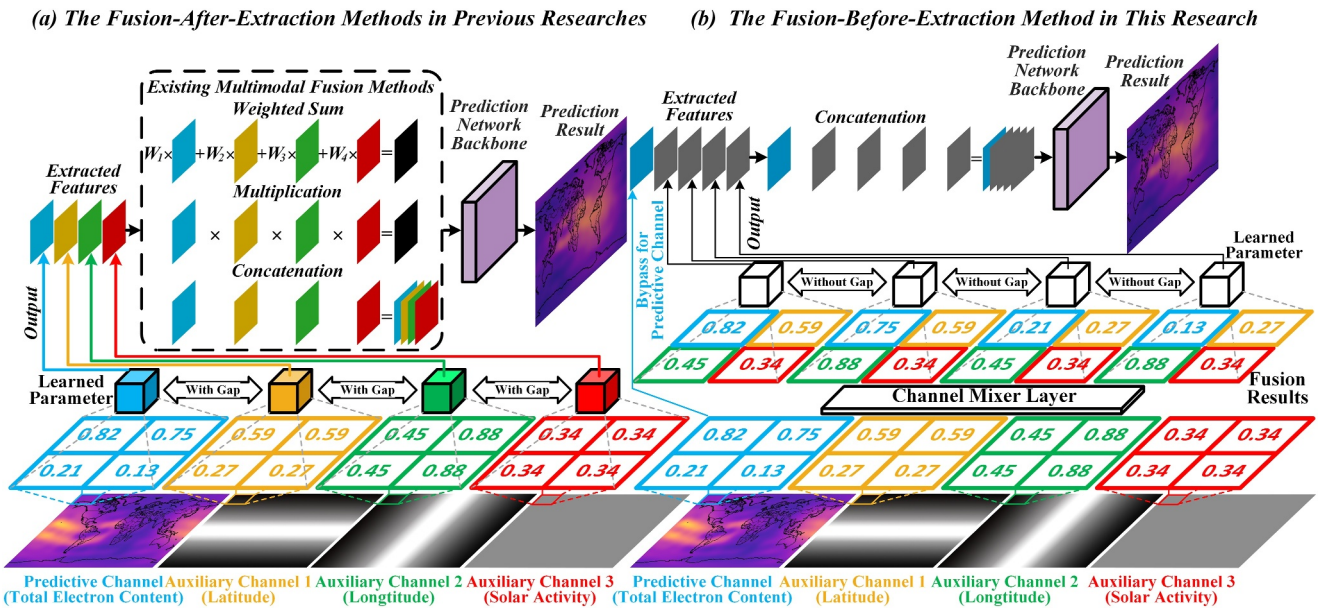

Figure 3. The flowchart of multimodal fusion methods for previous fusion-after-extraction approach (a) and our fusion-before-extraction approach (b).

Figure 4 shows the detailed implementation process of the proposed mixer layer. First, The input channels (C_k) are split along the height and width dimensions to get the individual data for all channels (SC). Then all of the individual data are concatenated along the channel dimension to get the pre-fusion channel (PC). Finally, grouping the data of pre-fusion channel and exchanging the positions of the specified adjacent groups, otherwise there are distribution differences along the width dimension for the fusion result. The fusion result of mixer layer is obtained after the size recovering for the fusion channel (FC). Here using the value assignment operation $\mathrm{FC}[n*i+k/n, n*j+k\%n] = C_k[i,j]$, where [ ] refers to height and width coordinates, n refers to the squared total channel number and % is the remainder operation, to realize mixer layer is easier to understand. But such an in-place operation can cause gradient interruption in the neural network, which is prohibited by the machine learning software (Bulò et al., 2018). In contrast, the proposed mixer layer only contains dimension

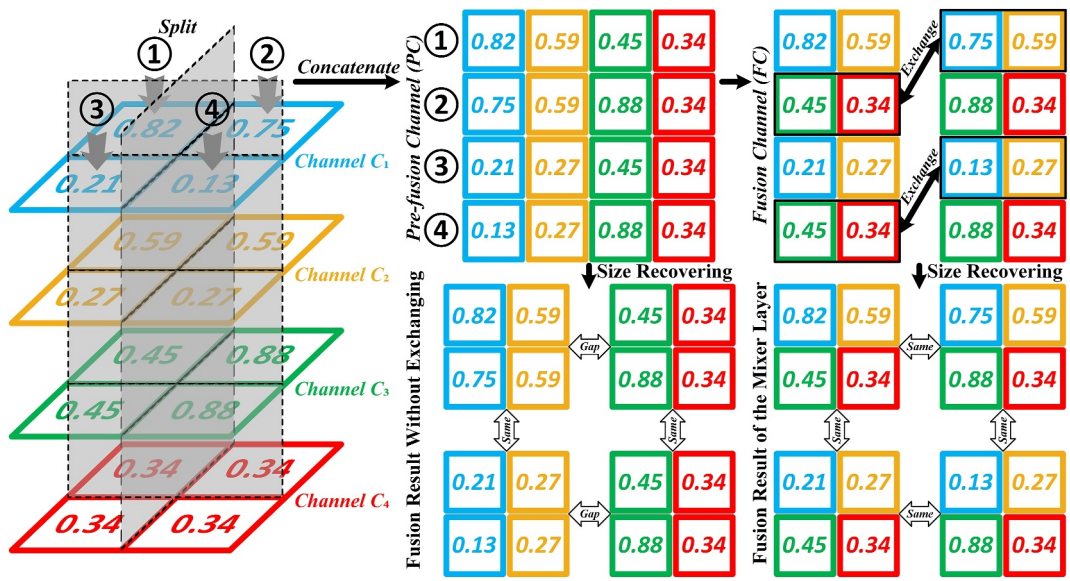

Figure 4. The detailed implementation flowchart of the proposed channel mixer layer. There are four procedures to get the final fusion result of the channel mixer layer which can eliminate the data distribution differences among channels: spatial dimension split, channel dimension concatenation, adjacent data exchange and size recovering.

transformations as shown by Formula 3, where all input and output data of each procedure are real values $\mathbb{R}$ with the size of positive integers $H, W, C \in \mathbb{Z}^+$ for height, width and channel dimensions respectively.

$$\begin{cases} \mathrm{SC} \in \mathbb{R}^{1,H,1,W,C} = \mathrm{Split}\left(\mathrm{input} \in \mathbb{R}^{H,W,C}\right) \\ \mathrm{PC} \in \mathbb{R}^{1,1,H*W*C} = \mathrm{Concatenate}(\mathrm{SC}) \\ \mathrm{FC} \in \mathbb{R}^{H*\sqrt{C},\sqrt{C},W/\sqrt{C},\sqrt{C}} = \mathrm{Exchange}(\mathrm{PC}) \\ \mathrm{output} \in \mathbb{R}^{H,W,C} = \mathrm{Recover}(\mathrm{FC}) \end{cases} \tag{3}$$

4. Result

In order to make the performance comparisons between models fair and persuasive, all RNN backbones are set as the same four stacked cell layers, there are 128 recurrent cells for each layer. All CNN backbones consist of the same 4-8-4 block layers for encoder-hidden-decoder respectively. There are more than 100 layers with the trainable parameters including the linear, normalization and convolution layers. All of the hyper-parameters such as filter size are set by default without changing. In terms of the training strategy, all models are trained in the same environment. The training data set is traversed the same epoch = 10 times repeatedly during the training phase. It takes about one day for the training of each model. After each training epoch is completed, the training data set is randomly shuffled meanwhile the learning rate is optimized by the scheduler automatically. The performance of different trained models in the test data set is evaluated from both quantitative and qualitative aspects which is further illustrated in the following subsections.

4.1. Quantitative Evaluation

In order to evaluate the prediction accuracy of different models more accurately, two routine metrics for the machine learning prediction are adopted in this research (Nigusie et al., 2024): Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The difference between them is that MSE is a dimensionless indicator without physical meaning whereas MAE and RMSE has the same unit as the input data that is TEC unit (TECU). These metrics can be calculated by Formula 4, where n refers to the total number of the data samples, $\hat{y}_i$ and y_i are the prediction result and the ground truth, respectively.

$$\begin{cases} \mathrm{MSE} = \dfrac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \\ \mathrm{MAE} = \dfrac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \\ \mathrm{RMSE} = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \end{cases} \tag{4}$$

Table 2 shows the quantitative performance evaluation including the computational complexity and the prediction error for the different network backbones with the different multimodal fusion methods that fuses the different input channel quantity (C_{in}). The indicators for computational complexity consist of the trainable parameter number (Param.), floating-point operations per second (Flops) and the inference speed in a unit of frame per second (fps). As for the indicators for the prediction error, the 4-day averaged MSE is calculated for the whole test data set then the zero-near result is scaled 100 times larger, whereas the 4-day averaged MAE is evaluated in the low solar activity (LSA) and high solar activity (HSA) periods, respectively. A model with more trainable parameters always has a higher accuracy so that the computational complexity of the different models is also shown in Table 2. Compared with the recurrent-free models like SimVP and TAU, the recurrent-based models always has the fewer parameters due to the parameter sharing at the different prediction time but the too small parameter number may also make the model unable to converge leading to under-fitting, for this reason, the performance of MAU backbone with the mixer layer is not listed in Table 2. However, the iterative prediction of recurrent structure also increases the computational complexity (Flops) by more than 20 times compared with that of recurrent-free models with the similar parameter number. Due to the high computational complexity, recurrent-

Table 2
The Quantitative Performance Comparison for the Different Network Backbones, Multimodal Fusion Methods and Input Channel Number C_{in}

						MAE/TECU		
Fusion-(Backbone)	C_{in}	Param/M	Flops/G	Speed/fps	MSE	LSA	HSA	RMSE (Gain)
(LSTM)	1	283.6	5,035	5	374.11	2.16	5.18	3.76 (+0%)
(GRU)	1	212.7	3,786	7	361.36	1.97	5.06	3.69 (+0%)
(E3D-LSTM)	1	53.15	301	22	173.51	1.11	3.54	2.56 (+0%)
(MAU)	1	4.72	67.13	242	153.97	1.10	3.28	2.41 (+0%)
Concatenate-	4	4.72	67.19	235	152.55	1.12	3.27	2.40 (+0.4%)
Concatenate-	9	4.72	67.28	215	139.90	1.08	3.11	2.30 (+4.6%)
Concatenate-	16	4.72	67.42	188	120.78	1.05	2.86	2.14 (+11.2%)
(PredRNN)	1	23.84	456	65	152.59	1.10	3.26	2.40 (+0%)
Action-	4	24.16	461	64	198.29	1.25	3.82	2.74 (−14.2%)
Concatenate-	4	24.92	476	62	172.22	1.21	3.49	2.55 (−6.3%)
Mixer-	**4**	**25.27**	**483**	**61**	**160.18**	**1.15**	**3.35**	**2.46 (−2.5%)**
Action-	9	24.51	468	64	277.31	1.47	4.17	3.24 (−35.0%)
Concatenate-	9	26.71	511	58	157.43	1.16	3.37	2.44 (−1.7%)
Mixer-	**9**	**27.06**	**518**	**57**	**155.98**	**1.16**	**3.32**	**2.43 (−1.3%)**
Action-	16	25.06	482	61	403.47	2.06	5.51	3.91 (−62.9%)
Concatenate-	16	29.22	558	53	138.65	1.11	3.10	2.29 (+4.6%)
Mixer-	**16**	**29.57**	**565**	**52**	**133.43**	**1.05**	**2.98**	**2.24 (+6.7%)**
(TAU)	1	48.24	22.54	341	151.06	1.09	3.26	2.39 (+0%)
Concatenate-	4	48.24	22.61	331	150.04	1.12	3.23	2.38 (+0.4%)
Mixer-	**4**	**48.24**	**22.61**	**326**	**148.83**	**1.11**	**3.20**	**2.37 (+0.8%)**
Concatenate-	9	48.24	22.83	302	142.98	1.04	3.18	2.33 (+2.5%)
Mixer-	**9**	**48.24**	**22.83**	**296**	**139.03**	**1.05**	**3.11**	**2.29 (+4.2%)**
Concatenate-	16	48.25	23.08	264	126.33	1.02	2.96	2.19 (+8.4%)
Mixer-	**16**	**48.25**	**23.11**	**254**	**121.54**	**0.99**	**2.89**	**2.14 (+10.5%)**
(SimVP)	1	50.53	23.32	335	149.51	1.05	3.24	2.38 (+0%)
Concatenate-	4	50.53	23.43	327	153.37	1.12	3.29	2.41 (−1.3%)
Mixer-	**4**	**50.53**	**23.46**	**322**	**152.96**	**1.11**	**3.28**	**2.40 (−0.8%)**
Concatenate-	9	50.53	23.61	298	145.15	1.08	3.19	2.34 (+1.7%)
Mixer-	**9**	**50.53**	**23.64**	**291**	**141.52**	**1.05**	**3.14**	**2.31 (+2.9%)**
Concatenate-	16	50.54	23.86	253	125.16	1.00	2.94	2.18 (+8.4%)
Mixer-	**16**	**50.54**	**23.90**	**241**	**120.74**	**0.98**	**2.90**	**2.13 (+10.5%)**
(ConvLSTM)	1	15.08	223	136	194.81	1.25	3.75	2.71 (+0%)
Concatenate-	4	15.70	235	132	199.78	1.28	3.79	2.75 (−1.5%)
Mixer-	**4**	**15.90**	**239**	**130**	**174.78**	**1.10**	**3.51**	**2.57 (+5.2%)**
Concatenate-	9	16.72	255	121	166.25	1.12	3.37	2.51 (+7.4%)
Mixer-	**9**	**16.93**	**258**	**118**	**165.87**	**1.09**	**3.38**	**2.50 (+7.8%)**
Concatenate-	16	18.16	282	105	147.21	1.08	3.20	2.36 (+12.9%)
Mixer-	**16**	**18.36**	**286**	**101**	**135.12**	**1.02**	**3.03**	**2.26 (+16.6%)**
(PredRNNv2)	1	23.86	458	63	152.01	1.07	3.25	2.40 (+0%)
Action-	4	24.18	463	62	167.69	1.15	3.34	2.52 (−5.0%)
Concatenate-	4	24.93	479	60	156.59	1.10	3.35	2.43 (−1.3%)

Table 2
Continued

Fusion-(Backbone)	C_{in}	Param/M	Flops/G	Speed/fps	MSE	MAE/TECU LSA	MAE/TECU HSA	RMSE (Gain)
Mixer-	**4**	**25.29**	**486**	**59**	**153.11**	**1.11**	**3.31**	**2.41 (−0.4%)**
Action-	9	24.53	470	61	237.44	1.36	3.92	2.99 (−24.6%)
Concatenate-	9	26.72	513	56	151.19	1.12	3.25	2.39 (+0.4%)
Mixer-	**9**	**27.08**	**520**	**54**	**150.28**	**1.10**	**3.23**	**2.38 (+0.8%)**
Action-	16	25.08	485	59	318.57	1.67	4.46	3.47 (−44.6%)
Concatenate-	16	29.23	561	51	122.86	1.01	2.88	2.16 (10.0%)
Mixer-	**16**	**29.59**	**568**	**51**	**118.83**	**0.99**	**2.83**	**2.12 (+11.7%)**
(MIM)	1	38.20	711	43	158.97	1.12	3.32	2.45 (+0%)
Concatenate-	4	39.28	731	42	158.27	1.11	3.32	2.45 (+0%)
Mixer-	**4**	**39.63**	**738**	**41**	**158.76**	**1.14**	**3.30**	**2.45 (+0%)**
Concatenate-	9	41.07	766	38	148.57	1.09	3.19	2.37 (+3.3%)
Mixer-	**9**	**41.42**	**773**	**37**	**145.33**	**1.08**	**3.18**	**2.34 (4.5%)**
Concatenate-	16	43.58	814	34	126.37	1.01	2.93	2.19 (+10.6%)
Mixer-	**16**	**43.93**	**821**	**32**	**118.72**	**0.98**	**2.90**	**2.12 (+13.5%)**
(PredRNN++)	1	38.58	675	46	153.12	1.05	3.26	2.41 (+0%)
Concatenate-	4	39.66	695	44	150.21	1.05	3.24	2.38 (+1.2%)
Mixer-	**4**	**40.01**	**702**	**43**	**148.83**	**1.10**	**3.21**	**2.37 (+1.7%)**
Concatenate-	9	41.45	730	41	144.20	1.11	3.16	2.34 (+2.9%)
Mixer-	**9**	**41.81**	**737**	**40**	**141.22**	**1.05**	**3.12**	**2.31 (+4.1%)**
Concatenate-	16	43.96	778	35	118.45	0.97	2.74	2.12 (+12.0%)
Mixer-	**16**	**44.32**	**785**	**34**	**111.68**	**0.94**	**2.63**	**2.05 (+14.9%)**

Note. The computational complexity is evaluated by parameter number (Param), Floating-point Operations Per Second (Flops) and inference speed in frame per second (fps). The prediction accuracy is evaluated by Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) during low/high solar activity (LSA/HSA) periods.

based models is slower than the recurrent-free models but all models are real-time with the inference speed of at least 40 fps when the batch = 1 sequence is input into the trained models serially on one NVIDIA A100 GPU device.

From the crosswise comparisons for the different network backbones in Table 2, it is obvious that the prediction accuracy of the spatiotemporal predictive models outperform the temporal predictive models such as LSTM and GRU. Moreover, the prediction error of ConvLSTM model is larger than other spatiotemporal predictive models proposed recently. From the longitudinal comparisons for the same network backbone in Table 2, as the number of input channel (C_{in}) increases, the prediction error of both concatenation and mixer multimodal fusion methods decreases, but that of the arithmetic multimodal fusion methods such as the weighted sum and multiplication (action-conditioned) increases on the contrary. The possible explanation is that these arithmetic methods essentially compress the number of the input channels to match the channel number of the TEC maps so that there is information loss in this procedure. In terms of the fusion methods without channel compression, compared with the graphic prediction when only global TEC map is input, the prediction error in MSE for both concatenation and mixer is almost same when four channels including the global TEC map and the spatial coordinates are input, possibly because the latitude and longitude channels cannot provide extra information due to the spatial learning ability that these backbones already have. The data distribution differences of concatenation method brings disadvantages in the learning procedure so that its performance is worse than that of mixer layer. Meanwhile, the performance of mixer layer has the lowest prediction error evaluated by MSE with a maximum improvement of about 15% (ConvLSTM and PredRNN++) with the 16-channel multimodal fusion prediction compared with that of one channel graphic prediction. During the LSA and HSA periods, the best accuracy of PredRNN++ backbone

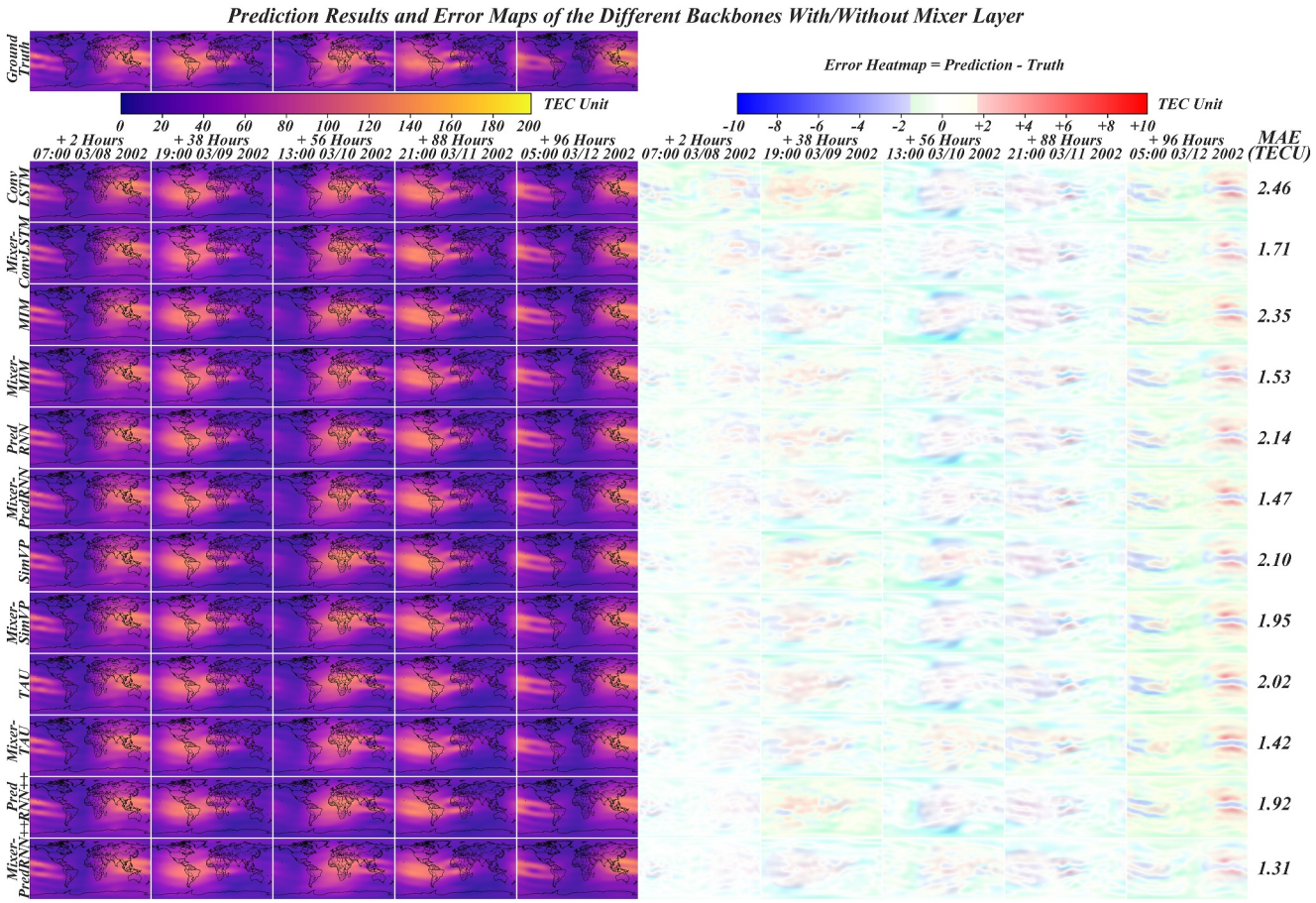

Figure 5. The qualitative performance comparison of prediction results (left) and error heatmaps (right) for the 1-channel graphic prediction and 16-channel mixer layer embedded multimodal fusion prediction of different model backbones.

with mixer layer reaches 0.94 and 2.63 TECU evaluated by MAE, respectively. All of the above evidence suggests that the our mixer layer has the best machine reasoning ability compared to the traditional multimodal fusion methods.

4.2. Qualitative Evaluation

Figure 5 shows the global TEC prediction examples in 96 hr of the different model backbones with/without the mixer multimodal fusion method during the high solar activity period from UT7:00, DOY067 to UT5:00, DOY072 in 2002. The left part in Figure 5 presents the results for the graphic prediction when inputting one channel of global TEC map and the multimodal fusion prediction using mixer layer when inputting $C_{\text{in}} = 16$ channels, respectively. Compared with the ground truth, it is obvious that the prediction result of multimodal fusion is closer to the ground truth, which is also proved by the error heatmaps (prediction − truth) as shown in the right part of Figure 5 with MAE error annotations more intuitively. With the prediction time passes by, the error between prediction results and ground truth becomes larger and larger for all models as shown by the thicker color in the error heatmaps. Different from other spatiotemporal sequence data sets, global TEC maps of two continuous days almost remain the same when the solar and geomagnetic conditions are similar. Although the prediction accuracy of different models during the solar and geomagnetic storms varies a lot as the detailed demonstration in Section 5, considering the low occurrence of these occasional events, the overall prediction results of different models are close in Figure 5.

Figure 6 is plotted to clarify the performance degradation of the different network backbones at each prediction timestamp ($2 \le T \le 96$ hours) evaluated by MAE metric, the solid and dashed lines represent the test data -set grouped during high and low solar activity periods respectively. Compared with the graphic prediction (red line)

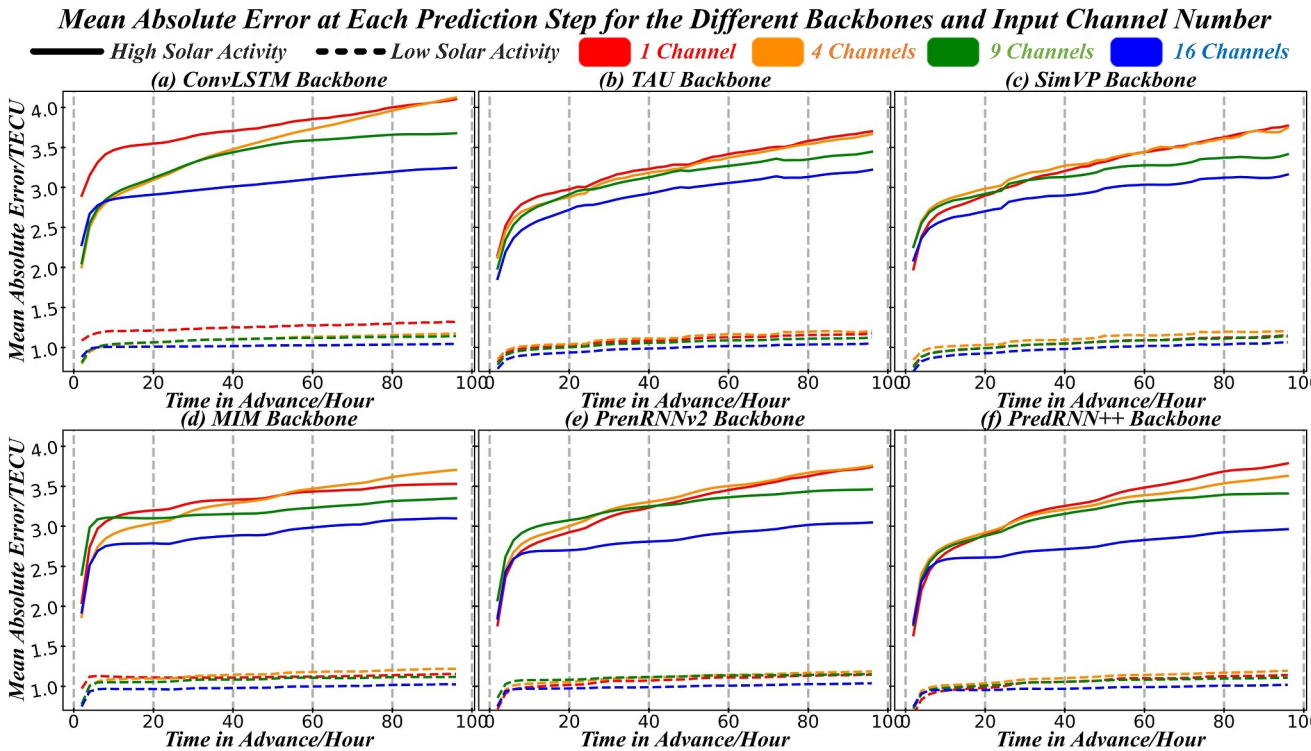

Figure 6. The accuracy degradation of the 1-channel graphic prediction and the 4/9/16-channel mixer layer embedded multimodal fusion prediction for six network backbones at each prediction timestamp during low/high solar activity periods.

with only previous global TEC maps input, the prediction accuracy of the machine reasoning prediction with four, nine and 16 channels (orange, green and blue lines) input into network backbones based on the mixer multimodal fusion method increases progressively. Different from previous researches (Ren et al., 2023) that aims to find out the external factor with the largest accuracy improvement contribution for a certain dependence, the framework proposed in this research makes it possible to input all relative auxiliary channels simultaneously so that models can make a more comprehensive decision by themselves. The mixer-PredRNN++ model trained with 16 input channels has the lowest MAE of 0.88/2.39, 0.94/2.59, 0.96/2.73 and 0.97/2.84 TECU during low/high solar activity period for the first, second, third and fourth day in advance respectively. For the features that all models share in common, the degradation of prediction accuracy turns from rapidly to slowly at about the tenth hour in advance.

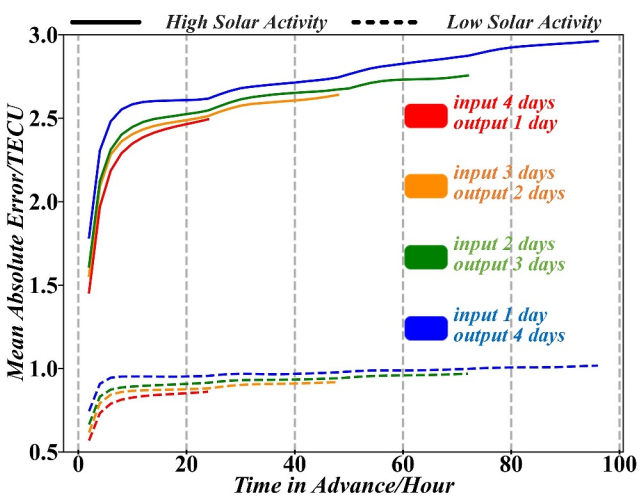

Figure 7. The prediction performance of the 16-channel mixer-PredRNN++ model with the different sequence length ratios of input and output.

The prediction accuracy of 16-channel mixer-PredRNN++ model that is trained with the different sequence length of input and output is shown as Figure 7. The total sequence length derived by the sliding window is fixed as 5 days to maintain the same data set, the portion ratios of the input and output sequence length are set as 4:1, 3:2, 2:3 and 1:4 as shown by the red, orange, green and blue curves respectively. It is obvious that the longer input length and shorter output length can lead to the higher prediction accuracy. Considering that the prediction error curves for different portion ratios only differ in quantity, the 1-day input and 4-day output strategy is adopted by default to be comparable with the similar researches that always predict more than future 3 days (Xia et al., 2022).

5. Discussion

The external dependence of global TEC and its prediction is discussed using mixer-PredRNN++ model with 16-channel input because it has the best performance as shown in last section. The detailed information is illustrated

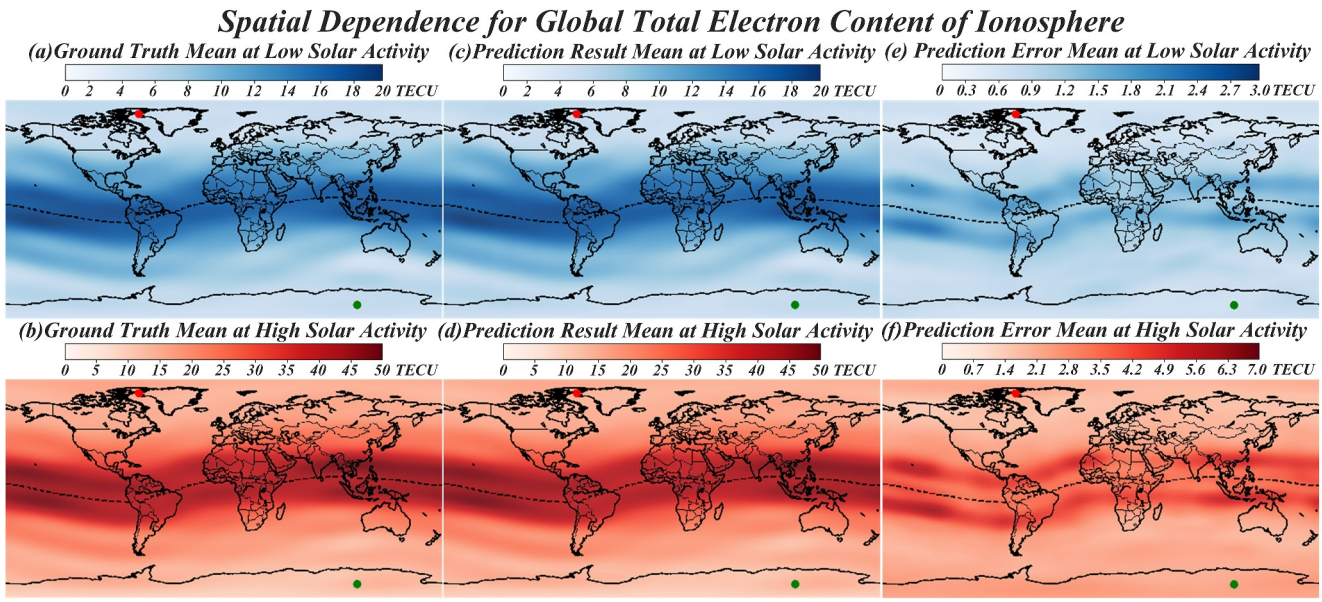

Figure 8. The spatial dependence of ionospheric total electron content during low (upper) and high (lower) solar activity. The north/south poles and equator in geomagnetic coordinate are annotated by the green/red dots and dashed line.

as the following subsections from the aspects of spatial, seasonal, diurnal, solar and geomagnetic activity dependence, respectively.

5.1. Spatial Dependence

Figure 8a, 8c and 8b, 8d show the spatial distribution of the averaged TEC observation ground truth (model prediction result) with the peak values of 16.85 (16.93) and 49.63 (49.14) TECU during the low and high solar activity periods respectively. Figures 8e and 8f show the spatial distribution of the averaged prediction error of the 16-channel mixer-PredRNN++ model evaluated by MAE metric with the peak values of 2.37 and 6.34 TECU during the low and high solar activity periods respectively. The north/south poles and equator in geomagnetic coordinate are also annotated by the green/red dots and dashed line in Figure 8. It is obvious that the spatial distribution of both observation ground truth and prediction error are symmetric about the geomagnetic equator, which is regarded as the evidence of the inter-hemispheric geomagnetic conjugate mechanism. Due to the Earth's rotation, the spatial distribution of global TEC is almost identical at the same latitude but different longitudes in geomagnetic coordinate. Notably, due to the lack of ground receiver, there is larger systematic error for the observation ground truth of about 5 TECU over the oceanic region (J. Chen et al., 2020).

5.2. Seasonal Dependence

Figure 9a, 9c and 9b, 9d show the seasonal dependence of the averaged TEC observation ground truth (model prediction result) at each latitude and day of year during the low and high solar activity periods respectively. Figures 9e and 9f show the seasonal dependence of the averaged TEC prediction error of the 16-channel mixer-PredRNN++ model at each latitude and day of year evaluated by MAE metric during the low and high solar activity periods respectively. It is obvious that both observation truth and prediction error reach to maximum at March and September equinox for the low latitude region, meanwhile, TEC reaches to maximum at June (December) solstice for the mid/high latitude of Northern (Southern) Hemisphere. The variation of solar zenith angle leads to the seasonal dependence of TEC.

5.3. Diurnal Dependence

Figure 10a, 10c and 10b, 10d show the diurnal dependence of the averaged TEC observation ground truth (model prediction result) at each universal time and longitude during the low and high solar activity periods respectively. Figures 10e and 10f show the diurnal dependence of the averaged TEC prediction error of the 16-channel mixer-PredRNN++ model at each universal time and longitude evaluated by MAE metric during the low and high solar

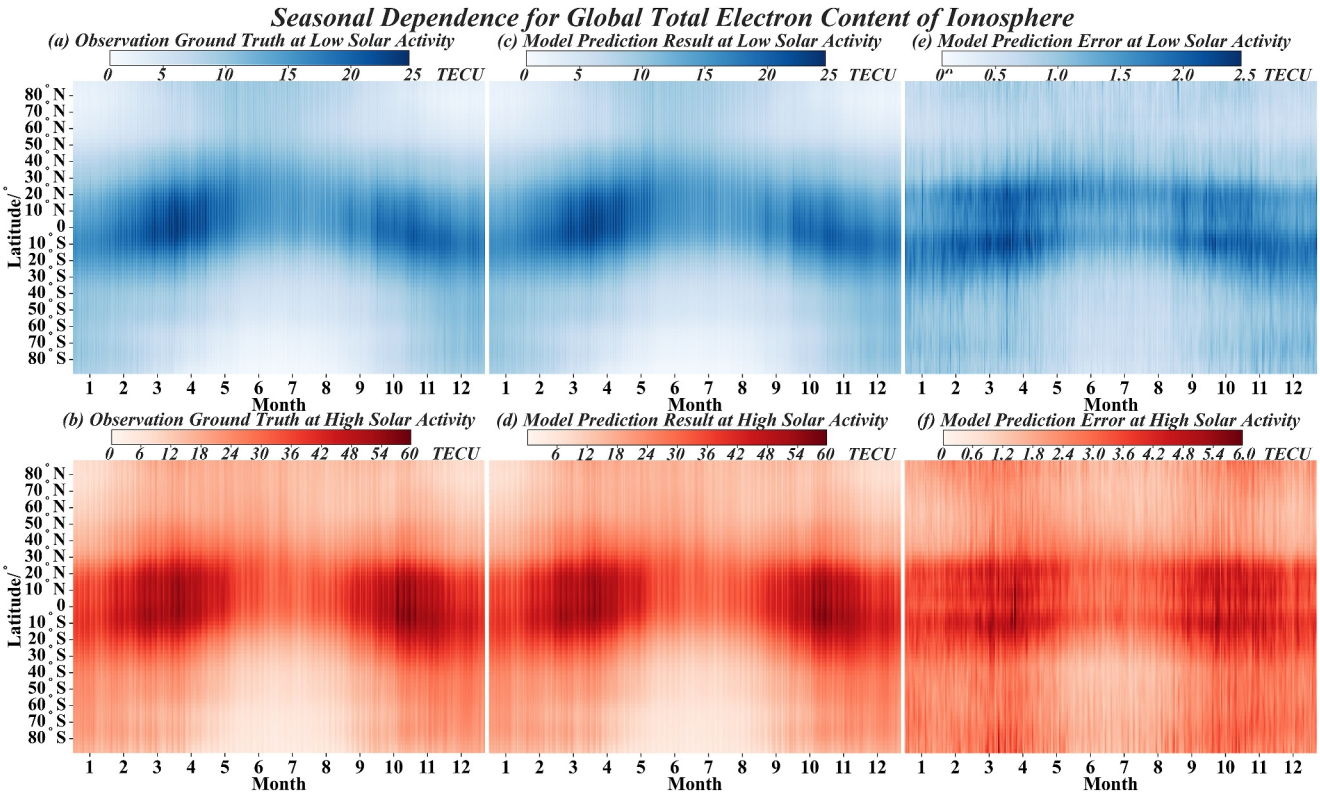

Figure 9. The seasonal dependence of ionospheric total electron content during low (upper) and high (lower) solar activity.

activity periods respectively. It is obvious that the local time (LT) at each longitude can be divided into diurnal or nocturnal time. TEC reaches to minimum at midnight LT 0 and maximum at midday LT 12. The neutral particles of upper atmosphere are ionized by the energy of sunlight, which is the most intense at the subsolar point leading to this diurnal dependence. Meanwhile, the TEC is higher at odd number hours during high solar activity period as shown by the horizontal stripes with the thicker color in three bottom subplots of Figure 10, that is because the observation before 2014 is only carried at odd number hours of universal time. The lack of observation TEC data at even number hours during higher solar activity period from 2000 to 2003 results in this difference.

5.4. Solar Activity Dependence

The more accurate prediction when more auxiliary channel input is only a necessary but insufficient condition for the assertion that the proposed mixer layer has the machine reasoning ability because the higher computational complexity may also lead to the accuracy improvement. The machine reasoning ability requires the model to correctly mine physical laws from data. Although aforementioned chapters qualitatively demonstrate that the model with mixer layer can learn the positive correlation between global TEC and solar activity, the quantitative experiment as shown in Figure 11 is still needed to prove the effectiveness of the proposed multimodal fusion method. The dependence analysis is illustrated by the linear regression fitting $\hat{y}_i$ and correlation coefficient R^2 as shown by Formula 5.

$$\begin{cases} \hat{y}_i = kx_i + b \\ R^2 = 1 - \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \end{cases} \tag{5}$$

The vertical and horizontal coordinates of each scatter point (x_i, y_i) in Figure 11a represent the daily averaged observation ground truth of global mean TEC and solar activity $F_{10.7}$ index (left) or sunspot number (right) from 1997 to 2022. Regression lines L_1 and L_2 are drawn based on the least squares method where the slope

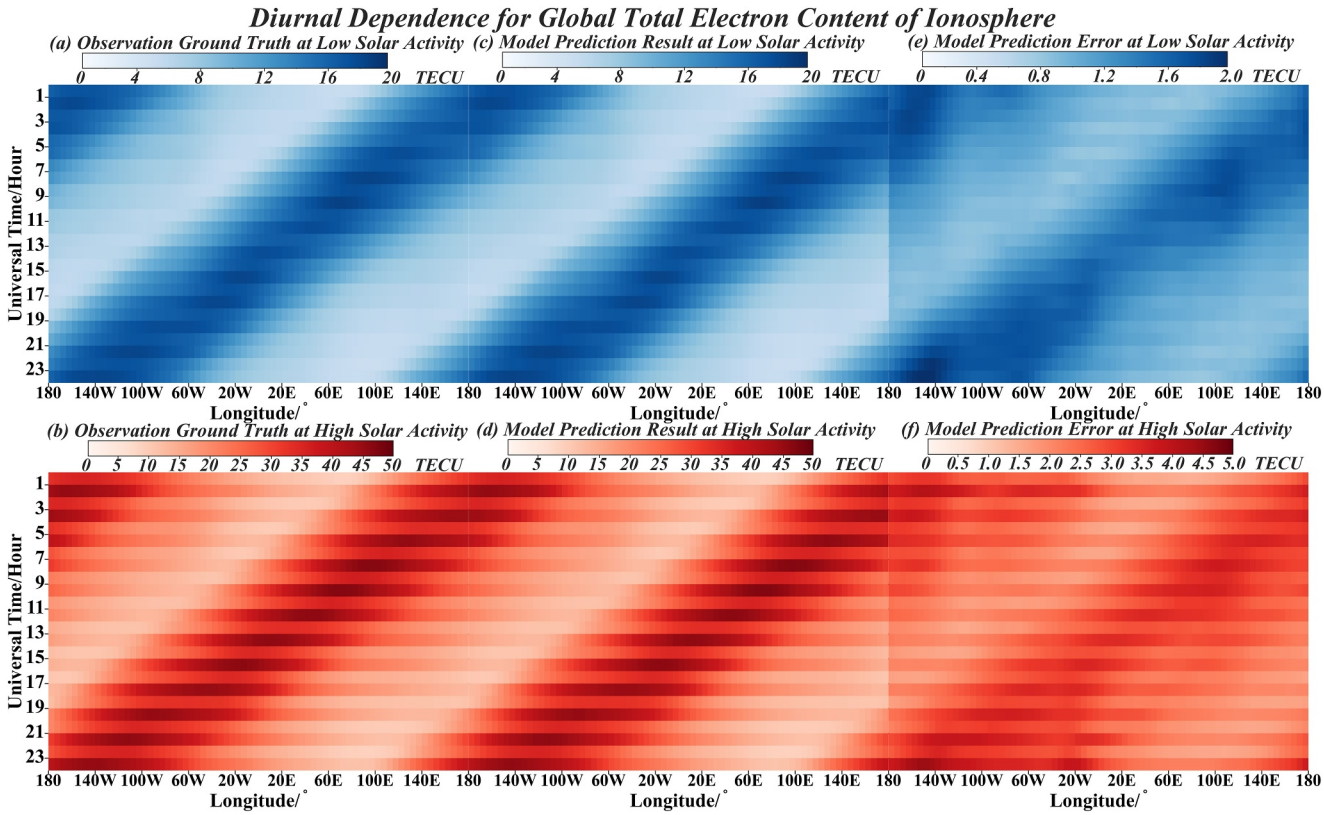

Figure 10. The diurnal dependence of ionospheric total electron content during low (upper) and high (lower) solar activity.

$k = \Delta y_i / \Delta x_i$ refers to the TEC variation caused by the solar activity variation. $R^2 = 0.78$ (0.69) are regarded as the ground truth correlation coefficients of physical laws between global mean TEC and solar activity $F_{10.7}$ (sunspot) index respectively. Figure 11b shows how the 9-channel mixer-PredRNN++ (black) and concatenate-PredRNN++ (orange) model predictions respond to the manually decreased/increased solar activity indices input. The red cross on vertical axis $\bar{y}_i = 14.93$ and Horizontal axis origin $\bar{x}_i = 57.3$ (103.3) refers to the average TEC and solar activity $F_{10.7}$ (sunspot) index of all points in Figure 11a. The scatter distribution in Figure 11a becomes more divergent during higher solar activity, indicating that the correlation is less significant. This also explains why the increasing speed of the model prediction results in Figure 11b decreases as the increase of solar activity. Compared to concatenate-PredRNN++ model L_4, the regression line slope of mixer-PredRNN++ model L_3 during lower solar activity is closer to observation ground truth. The correlation coefficients $R^2 = 0.61$ (0.53) between TEC and solar activity $F_{10.7}$ (sunspot) index calculated by regression line L_3 of mixer-PredRNN++ model for all scatter points in Figure 11a are also closer to ground truth compared with that $R^2 = 0.25$ (0.48) of concatenate-PredRNN++ model.

5.5. Geomagnetic Activity Dependence

Geomagnetic storm brings damage to human society leading to the electrical system disruption, satellite signal scintillation and larger navigation error. Geomagnetic quiet period is characterized as the near zero Dst and Ap geomagnetic activity indices (Ren et al., 2024). Figure 12a shows how Dst (blue solid line) and Ap (purple solid line) indices fluctuated during eight geomagnetic storm events in the past two decades. Figure 12b shows the global mean TEC of observation ground truth (black solid line) or prediction results of IRI2020 model (cyan dotted line), 1-channel graphic PredRNN++ model (green dashed line), 16-channel concatenate-PredRNN++ multimodal fusion model (orange dashed line) and 16-channel mixer-PredRNN++ multimodal fusion model (red dashed line). It is obvious that during both low (2017 of event 8) and high (others) solar activity years, the global mean TEC of IRI2020 simulation is lower than that of observation ground truth due to the exclusion of

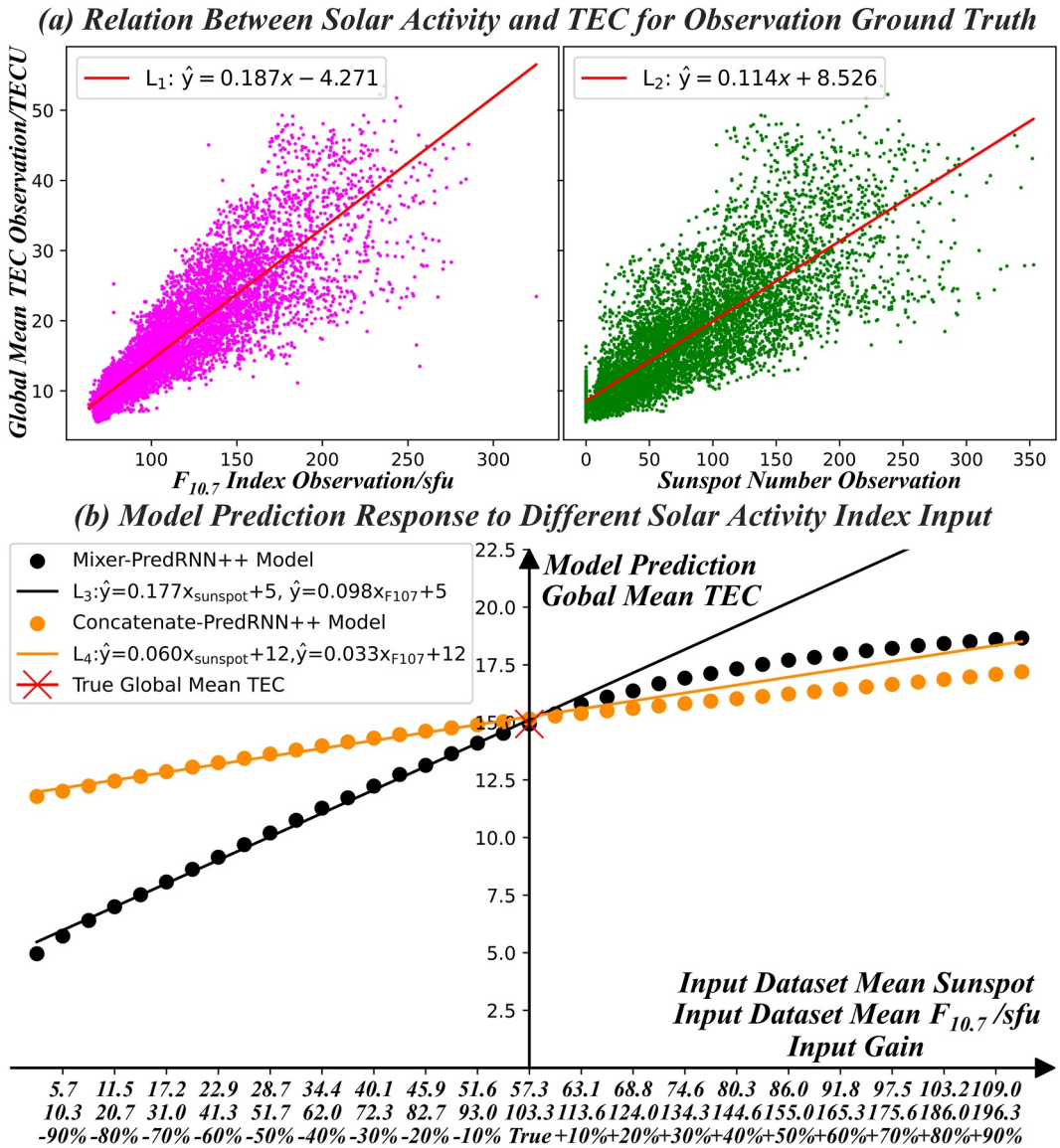

Figure 11. The machine reasoning performance from solar activity input to TEC prediction by different multimodal fusion methods. The response regression line of proposed mixer layer is closer to observation ground truth.

plasmaspheric TEC contribution (Bilitza et al., 2022). Since unable to learn the geomagnetic activity feature, the 4-day TEC prediction output of 1-channel PredRNN++ model is roughly equivalent to the 1-day global mean TEC input, which cannot respond to the geomagnetic storm. Although both concatenate-PredRNN++ and mixer-PredRNN++ models take geomagnetic activity into consideration, the prediction result of mixer-PredRNN++ model proposed in this research is closer to ground truth.

6. Conclusion

In this research, a new multimodal fusion method named as channel mixer layer is proposed which can be embedded into the existing advanced spatiotemporal sequence predictive models to improve the global TEC prediction accuracy by inputting auxiliary channel data such as time and coordinates. The working principle of channel mixer layer is eliminating the data distribution differences among different concatenated channels so that the learned feature of the same model parameter can remain consistency while sliding across different channels according to the parameter sharing of neural network. To achieve the best prediction performance, a new standard data set consist of the largest sequence quantity of 106,609 are released where the global TEC maps and their

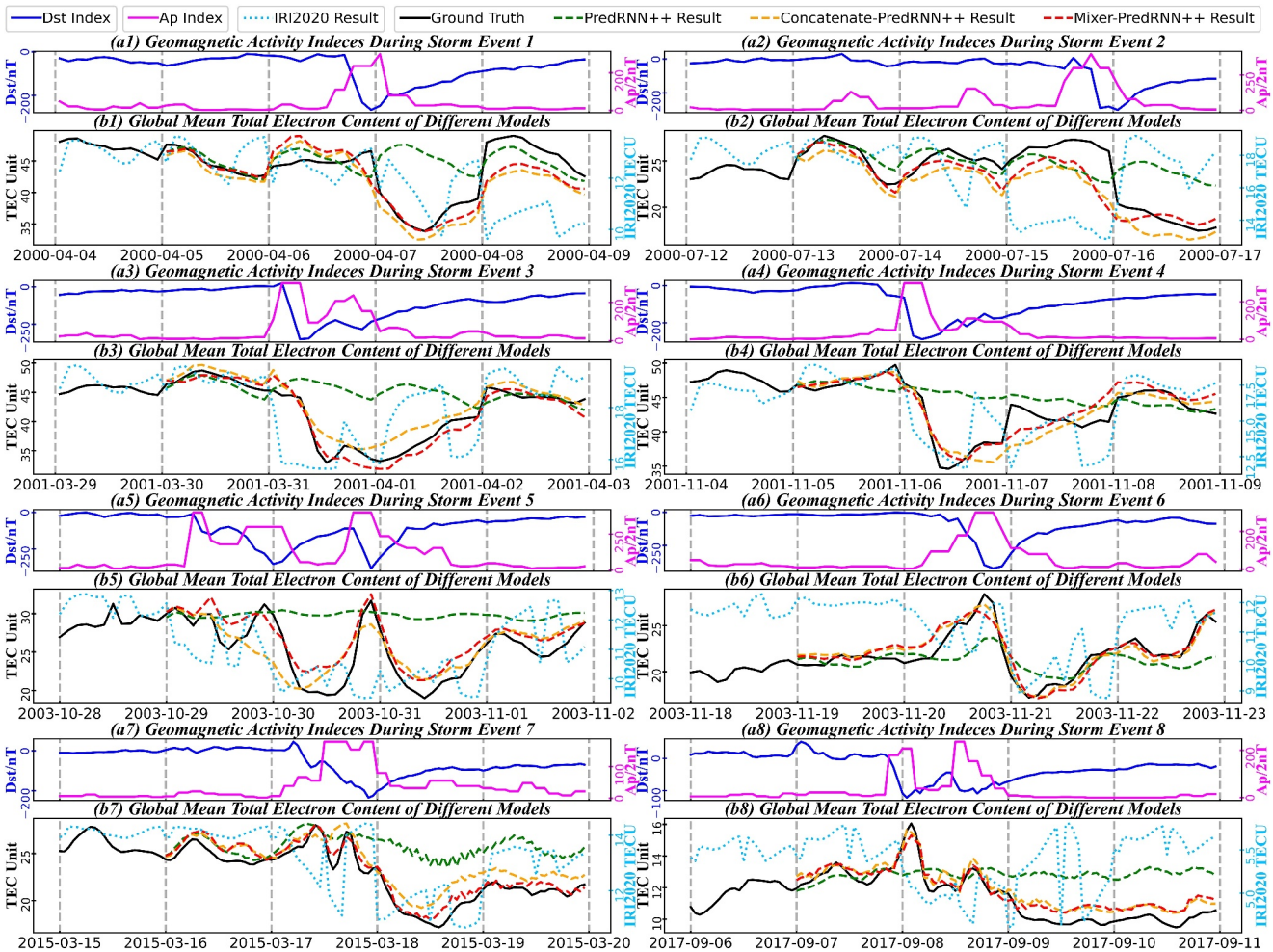

Figure 12. The panel (a) shows the temporal fluctuation of geomagnetic activity Dst (blue solid line) and Ap (purple solid line) indices. The panel (b) shows the global mean TEC of observation ground truth (black solid line) or prediction results of IRI2020 model (cyan dotted line), 1-channel graphic PredRNN++ model (green dashed line), 16-channel concatenate-PredRNN++ multimodal fusion model (orange dashed line) and our 16-channel mixer-PredRNN++ multimodal fusion model (red dashed line).

corresponding external factors are packaged together in the specified format of machine learning software for the high speed accessing. With the same data set and running environment, the experiment result suggests that the channel mixer layer embedded models that extract features after multimodal fusion outperforms other feature-level multimodal fusion methods that extract features before fusion such as concatenation and multiplication. Meanwhile, more input data channel and larger sequence length ratio of input to output also contribute to the prediction accuracy improvement. The best prediction performance of 16-channel mixer-PredRNN++ model when 1-day sequence input and 4-day sequence output haves the minimum mean absolute error of 0.94 and 2.63 TEC unit during low and high solar activity period with the real-time inference speed of 34 frames per second.

In this research, the external dependence of global TEC is illustrated from the spatial, seasonal, diurnal, solar and geomagnetic activity aspects. The movement of subsolar point and geomagnetic conjugate mechanism are the reasons for the spatial, seasonal and diurnal dependence of global TEC spatiotemporal distribution. To prove the machine reasoning ability of channel mixer layer embedded models that can correctly mine physical laws from data, a quantitative experiment that how model predictions respond to the manually decreased/increased solar activity input indices is conducted where results suggest that the regression line slopes and correlation coefficients between solar activity indices and global mean TEC predicted by mixer-PredRNN++ model is closer to observation ground truth compared to that of concatenate-PredRNN++ model. In terms of geomagnetic activity dependence, the graphic prediction results of original PredRNN++ model without auxiliary data input cannot

respond to the geomagnetic storm, whereas the multimodal fusion prediction results of proposed mixer-PredRNN++ model with auxiliary data input is closer to observation ground truth compared with other multimodal fusion methods.

In the future, the proposed prediction framework is expected to integrate into the TEC observation system, contributing to the economic and industrial development such as radio quality improvement and space weather forecasting.

Data Availability Statement

The whole program of this research is open-sourced at Github and Zenodo (P. Liu, 2024a). The IonoElectron data set proposed in this research for the Tensorflow and Pytorch machine learning software is available at Zenodo (P. Liu, 2024b), which is made by packaging the TEC map derived from Centre for Orbit Determination in Europe (CODE): http://www.aiub.unibe.ch/download/CODE, together with the auxiliary factors such as solar and geomagnetic activity indices derived from OMNI data set: https://spdf.gsfc.nasa.gov/pub/data/omni/low_res_omni/.

Acknowledgments
This work is supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant 22K21345, 21H04518, 24KJ0125 and 20H00197. This work is a part of JSPS Grant-in-Aid for International Leading Research (PBASE program). The authors thank the calculation hardware supported by A-KDK supercomputer at Research Institute for Sustainable Humanosphere, Kyoto University.

References

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Bilitza, D., Pezzopane, M., Truhlik, V., Altadill, D., Reinisch, B. W., & Pignalberi, A. (2022). The international reference ionosphere model: A review and description of an ionospheric benchmark. *Reviews of Geophysics*, *60*(4), e2022RG000792. https://doi.org/10.1029/2022RG000792

Bulò, S. R., Porzi, L., & Kontschieder, P. (2018). In-place activated batchnorm for memory-optimized training of dnns. In *2018 ieee/cvf conference on computer vision and pattern recognition* (pp. 5639–5647). https://doi.org/10.1109/CVPR.2018.00591

Chang, Z., Zhang, X., Wang, S., Ma, S., Ye, Y., Xinguang, X., & Gao, W. (2021). Mau: A motion-aware unit for video prediction and beyond. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 26950–26962). Curran Associates, Inc. Retrieved from. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2021/file/e25cfa90f04351958216f97e3efdabe9-Paper.pdf

Chen, J., Ren, X., Zhang, X., Zhang, J., & Huang, L. (2020). Assessment and validation of three ionospheric models (iri-2016, nequick2, and igs-gim) from 2002 to 2018. *Space Weather*, *18*(6), e2019SW002422. https://doi.org/10.1029/2019SW002422

Chen, Z., Liao, W., Li, H., Wang, J., Deng, X., & Hong, S. (2022). Prediction of global ionospheric tec based on deep learning. *Space Weather*, *20*(4), e2021SW002854. https://doi.org/10.1029/2021SW002854

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Nips 2014 workshop on deep learning, December 2014*.

Codrescu, M. V., Negrea, C., Fedrizzi, M., Fuller-Rowell, T. J., Dobin, A., Jakowsky, N., et al. (2012). A real-time run of the coupled thermosphere ionosphere plasmasphere electrodynamics (ctipe) model. *Space Weather*, *10*(2). https://doi.org/10.1029/2011SW000736

Dow, J. M., Neilan, R. E., & Rizos, C. (2009). The international gnss service in a changing landscape of global navigation satellite systems. *Journal of Geodesy*, *83*(3), 191–198. https://doi.org/10.1007/s00190-008-0300-3

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Fu, W., Yokoyama, T., Ssessanga, N., Yamamoto, M., & Liu, P. (2022). On using a double-thin-shell approach and tec perturbation component to sound night-time mid-latitude e–f coupling. *Earth Planets and Space*, *74*(1), 83. https://doi.org/10.1186/s40623-022-01639-w

Gao, Z., Tan, C., Wu, L., & Li, S. Z. (2022). Simvp: Simpler yet better video prediction. In *2022 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 3160–3170). https://doi.org/10.1109/CVPR52688.2022.00317

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Karatay, S., & Gul, S. E. (2023). Prediction of gps-tec on mw > 5 earthquake days using bayesian regularization backpropagation algorithm. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5. https://doi.org/10.1109/LGRS.2023.3262028

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Li, W., Zhu, H., Shi, S., Zhao, D., Shen, Y., & He, C. (2024). Modeling China's sichuan-yunnan's ionosphere based on multichannel woa-cnn-lstm algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–18. https://doi.org/10.1109/TGRS.2024.3403684

Liu, J., Chen, R., Wang, Z., & Zhang, H. (2011). Spherical cap harmonic model for mapping and predicting regional tec. *GPS Solutions*, *15*(2), 109–119. https://doi.org/10.1007/s10291-010-0174-8

Liu, L., Zou, S., Yao, Y., & Wang, Z. (2020). Forecasting global ionospheric tec using deep learning approach. *Space Weather*, *18*(11), e2020SW002501. https://doi.org/10.1029/2020SW002501

Liu, P. (2024a). DrHoisu/Channel_Mixer_Layer: Software for space weather journal paper. *Zenodo*. https://doi.org/10.5281/zenodo.14227169

Liu, P. (2024b). Iono_Electron dataset for multimodal fusion and spatiotemporal predictive learning. *Zenodo*. https://doi.org/10.5281/zenodo.13165939

Liu, P., Yokoyama, T., Fu, W., & Yamamoto, M. (2022). Statistical analysis of medium-scale traveling ionospheric disturbances over Japan based on deep learning instance segmentation. *Space Weather*, *20*(7), e2022SW003151. https://doi.org/10.1029/2022SW003151

Luo, H., Gong, Y., Chen, S., Yu, C., Yang, G., Yu, F., et al. (2023). Prediction of global ionospheric total electron content (tec) based on sam-convlstm model. *Space Weather*, *21*(12), e2023SW003707. https://doi.org/10.1029/2023SW003707

Nava, B., Coïsson, P., & Radicella, S. (2008). A new version of the nequick ionosphere electron density model. *Journal of Atmospheric and Solar-Terrestrial Physics*, *70*(15), 1856–1862. https://doi.org/10.1016/j.jastp.2008.01.015

Nigusie, A., Tebabal, A., & Galas, R. (2024). Modeling ionospheric tec using gradient boosting based and stacking machine learning techniques. *Space Weather*, *22*(3), e2023SW003821. https://doi.org/10.1029/2023SW003821

Perlongo, N. J., Ridley, A. J., Cnossen, I., & Wu, C. (2018). A year-long comparison of gps tec and global ionosphere-thermosphere models. *Journal of Geophysical Research: Space Physics*, *123*(2), 1410–1428. https://doi.org/10.1002/2017JA024411

Rawer, K., Bilitza, D., & Ramakrishnan, S. (1978). Goals and status of the international reference ionosphere. *Reviews of Geophysics*, *16*(2), 177–181. https://doi.org/10.1029/RG016i002p00177

Ren, X., Yang, P., Liu, H., Chen, J., & Liu, W. (2022). Deep learning for global ionospheric tec forecasting: Different approaches and validation. *Space Weather*, *20*(5), e2021SW003011. https://doi.org/10.1029/2021SW003011

Ren, X., Yang, P., Mei, D., Liu, H., Xu, G., & Dong, Y. (2023). Global ionospheric tec forecasting for geomagnetic storm time using a deep learning-based multi-model ensemble method. *Space Weather*, *21*(3), e2022SW003231. https://doi.org/10.1029/2022SW003231

Ren, X., Zhao, B., Ren, Z., Wang, Y., & Xiong, B. (2024). Deep learning-based prediction of global ionospheric tec during storm periods: Mixed cnn-bilstm method. *Space Weather*, *22*(7), e2024SW003877. https://doi.org/10.1029/2024SW003877

Ridley, A., Deng, Y., & Tóth, G. (2006). The global ionosphere–thermosphere model. *Journal of Atmospheric and Solar-Terrestrial Physics*, *68*(8), 839–864. https://doi.org/10.1016/j.jastp.2006.01.008

Ruwali, A., Kumar, A. J. S., Prakash, K. B., Sivavaraprasad, G., & Ratnam, D. V. (2021). Implementation of hybrid deep learning model (lstm-cnn) for ionospheric tec forecasting using gps data. *IEEE Geoscience and Remote Sensing Letters*, *18*(6), 1004–1008. https://doi.org/10.1109/LGRS.2020.2992633

Shi, S., Zhang, K., Wu, S., Shi, J., Hu, A., Wu, H., & Li, Y. (2022). An investigation of ionospheric tec prediction maps over China using bidirectional long short-term memory method. *Space Weather*, *20*(6), e2022SW003103. https://doi.org/10.1029/2022SW003103

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf

Srivani, I., Siva Vara Prasad, G., & Venkata Ratnam, D. (2019). A deep learning-based approach to forecast ionospheric delays for gps signals. *IEEE Geoscience and Remote Sensing Letters*, *16*(8), 1180–1184. https://doi.org/10.1109/LGRS.2019.2895112

Tan, C., Gao, Z., Wu, L., Xu, Y., Xia, J., Li, S., & Li, S. Z. (2023). Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *2023 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 18770–18782). https://doi.org/10.1109/CVPR52729.2023.01800

Tang, J., Xu, L., Wu, X., & Chen, K. (2024). A short-term forecasting method for ionospheric tec combining local attention mechanism and lstm model. *IEEE Geoscience and Remote Sensing Letters*, *21*, 1–5. https://doi.org/10.1109/LGRS.2024.3373457

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., et al. (2021). Mlp-mixer: An all-mlp architecture for vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 24261–24272). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2021/file/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Paper.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In I. Guyon (Ed.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, H., Liu, H., Yuan, J., Le, H., Shan, W., & Li, L. (2024). Maooa-residual-attention-biconvlstm: An automated deep learning framework for global tec map prediction. *Space Weather*, *22*(7), e2024SW003954. https://doi.org/10.1029/2024SW003954

Wang, Y., Gao, Z., Long, M., Wang, J., & Yu, P. S. (2018). PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 5123–5132). PMLR. Retrieved from https://proceedings.mlr.press/v80/wang18b.html

Wang, Y., Jiang, L., Yang, M.-H., Li, L.-J., Long, M., & Fei-Fei, L. (2019). Eidetic 3d LSTM: A model for video prediction and beyond. In *International conference on learning representations*. Retrieved from https://openreview.net/forum?id=B1lKS2AqtX

Wang, Y., Long, M., Wang, J., Gao, Z., & Yu, P. S. (2017). Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In I. Guyon (Ed.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/e5f6ad6ce374177eef023bf5d0c018b6-Paper.pdf

Wang, Y., Sun, F., Huang, W., He, F., & Tao, D. (2023). Channel exchanging networks for multimodal and multitask dense image prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 5481–5496. https://doi.org/10.1109/TPAMI.2022.3211086

Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P. S., & Long, M. (2023). Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 2208–2225. https://doi.org/10.1109/TPAMI.2022.3165153

Wang, Y., Zhang, J., Zhu, H., Long, M., Wang, J., & Yu, P. S. (2019). Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *2019 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 9146–9154). https://doi.org/10.1109/CVPR.2019.00937

Wang, Z., Zou, S., Sun, H., & Chen, Y. (2023). Forecast global ionospheric tec: Apply modified u-net on vista tec data set. *Space Weather*, *21*(8), e2023SW003494. https://doi.org/10.1029/2023SW003494

Xia, G., Zhang, F., Wang, C., & Zhou, C. (2022). Ed-convlstm: A novel global ionospheric total electron content medium-term forecast model. *Space Weather*, *20*(8), e2021SW002959. https://doi.org/10.1029/2021SW002959

Xiong, P., Zhai, D., Long, C., Zhou, H., Zhang, X., & Shen, X. (2021). Long short-term memory neural network for ionospheric total electron content forecasting over China. *Space Weather*, *19*(4), e2020SW002706. https://doi.org/10.1029/2020SW002706

Xu, C., Ding, M., & Tang, J. (2024). Prediction of gnss-based regional ionospheric tec using a multichannel convlstm with attention mechanism. *IEEE Geoscience and Remote Sensing Letters*, *21*, 1–5. https://doi.org/10.1109/LGRS.2024.3373445

Zhou, Y., Liu, J., Li, S., & Li, Q. (2024). Ionospheric tec prediction based on ensemble learning models. *Space Weather*, *22*(3), e2023SW003790. https://doi.org/10.1029/2023SW003790