Applied Problems of Nonlinear Oscillation and Wave Theory

Izvestiya Vysshikh Uchebnykh Zavedeniy. Applied Nonlinear Dynamics. 2026;34(2)

Article DOI: 10.18500/0869-6632-003207

A hybrid total electron content forecasting model based on autoencoders and classical machine learning algorithms

A. M. Appalonov✉, *Y. S. Maslennikova*

Kazan Federal University, Russia
E-mail: ✉artem309_97@mail.ru, yuliamsl@gmail.com
Received 28.11.2025, accepted 25.12.2025, available online 26.12.2025, published 31.03.2026

Abstract. *Purpose.* Development of a novel two-stage machine learning algorithm for total electron content (TEC) forecasting based on original TEC time series and influential external ionospheric parameters. *Methods.* Dimensionality reduction of the input data is performed using a standard fully-connected autoencoder to obtain compressed latent representations. These features are integrated with a set of external parameters: the critical frequency of the F2 layer (foF2), solar (F10.7) and geomagnetic (Kp) activity indices, and temporal descriptors (seasonal and diurnal information). The enriched dataset is used to train and evaluate several classical machine learning algorithms, including gradient boosting (CatBoost), with assessment based on RMSE and MAE metrics. *Results.* The CatBoost algorithm demonstrates superior predictive accuracy on the test dataset compared to other evaluated models. The proposed two-stage approach proves effective for extracting and utilizing key temporal dependencies for the regression task. *Conclusion.* The developed method provides accurate TEC prediction by combining neural network-based time series compression with modern ensemble algorithms, as confirmed by the computational experiment.

Keywords: ionosphere, total electronic content, TEC maps, solar activity, machine learning, neural networks, auto-coder, equatorial anomaly.

For citation: Appalonov AM, Maslennikova YS. A hybrid total electron content forecasting model based on autoencoders and classical machine learning algorithms. Izvestiya VUZ. Applied Nonlinear Dynamics. 2026;34(2): 286–298. DOI: 10.18500/0869-6632-003207

Introduction

The Earth's ionosphere, which is an ionized part of the upper atmosphere, plays a critical role in the propagation of radio waves and the functioning of modern technological systems. The key parameter integrally characterizing the state of the ionosphere is the total electronic content (TEC) [1]. TEC fluctuations have a significant impact on the accuracy and reliability of satellite navigation systems (such as GLONASS and GPS), communication systems and remote sensing of the Earth. Sudden TEC disturbances caused by solar and geomagnetic activity can lead to significant positioning errors, poor communication quality, and even disruptions in the operation of energy networks [2, 3]. In this regard, the task of developing accurate and reliable methods for predicting the state of the ionosphere, and TEC in particular, remains highly relevant for fundamental science and practical applications.

Traditional approaches to ionospheric modeling can be roughly divided into physical and empirical ones. Physical models based on solving complex systems of equations describing physico-chemical processes in the ionosphere require significant computational resources and often cannot keep up with the rapid changes in it [4]. Empirical models based on statistical analysis of large data archives are more efficient, but they may not accurately take into account the specifics of particular geophysical conditions. Both approaches face difficulties in describing nonlinear and non-stationary time series, such as the TEC measurements [5, 6].

Recently, machine learning (ML) has demonstrated significant potential in solving time series forecasting problems in geophysics. Algorithms such as the support vector machine (SVM), random forest, and gradient boosting have been successfully used to model TEC [7]. However, the effectiveness of these methods directly depends on the quality and representativeness of the feature space. The initial TEC time series are characterized by high dimensionality, the presence of noise and complex time dependencies, which makes it difficult to use them directly in classical MO algorithms. This necessitates the stage of data preprocessing and the extraction of informative features.

One of the powerful methods for reducing dimensionality and extracting hidden patterns in data is the autoencoders [8]. These neural networks are able to learn a compressed representation of the source data, filtering out noise and preserving the most significant dependencies. Unlike convolutional networks focused on spatial patterns, fully-connected autoencoders are well suited for working with time series, allowing them to effectively reduce their dimension without losing significant information. The latent representations obtained at their output form a compact and informative description of the dynamics of ionospheric processes.

In addition, the state of the ionosphere depends on many external factors. These include parameters characterizing solar activity (such as the F10.7 index), geomagnetic disturbance (Kp index), as well as key ionospheric characteristics, such as the critical frequency of the F2 layer (foF2). Diurnal and seasonal variations also contribute significantly to the variability of the TEC. Thus, combining latent representations of the TEC time series with relevant external features makes it possible to create a complex feature space that significantly increases the potential of forecasting models.

The purpose of this study is to develop a hybrid model for predicting the total electronic content of the ionosphere, combining deep learning methods to identify key features and classical machine learning algorithms for finite regression.

Tasks to be solved:

1) building and training a fully-connected autoencoder for TEC time series compression and latent feature extraction;
2) formation of an extended dataset by combining these features with external geophysical and temporal parameters;
3) comparative analysis of the effectiveness of various classical ML algorithms, including gradient boosting, on an expanded dataset.

1. Neural network architecture for data dimensionality reduction

1.1. Rationale for choosing a neural network architecture. To effectively solve the problem of predicting ionospheric parameters, the correct preparation of the feature space is a critical step. The initial data, including time series of maps of the total electronic content (TEC) and a series of values

of the critical frequency of the F2 layer (foF2), are characterized by high dimensionality, the presence of noise and complex nonlinear dependencies. The direct use of such data for classical machine learning algorithms often leads to overfitting and a decrease in the generalizing ability of models. In this regard, this paper uses an approach based on an autoencoder, a special neural network architecture designed to effectively reduce the dimensionality of data and extraction the most significant latent features.

The main task to be solved at this stage was the nonlinear compression of the original feature space to a compact vector representation with a dimension of 10 elements. This makes it possible to significantly reduce the computational complexity of subsequent analysis, filter out noise, and identify invariant representations that preserve key information about the dynamics of ionospheric processes.

1.2. The structure of the autoencoder. The proposed architecture of the autoencoder (Fig. 1) is a sequential model consisting of two symmetrical parts: an encoder and a decoder.

The input vector of the model is a concatenation of the «straightened» ($M\times N\to M\cdot N$) target TEC map and the corresponding stretched map of foF2 values constructed for the same time point. The use of this additional spatial information improves the quality of the reconstruction [9].

The main advantage of this approach lies in its ability to identify complex nonlinear dependencies in data and extract the most informative features automatically, without explicitly specifying the transformation rules. Unlike traditional dimensionality reduction methods such as principal component analysis, autoencoders are able to account for complex structural features of data, including hierarchical relationships between features.

Autoencoders are of particular value when working with spatiotemporal data, such as global maps of ionospheric parameters. In this case, they allow not only to effectively reduce the dimensionality of the data, but also to identify hidden patterns and anomalies that are difficult to detect using traditional methods.

The initial dataset was divided in a ratio of 80:20 for the purposes of training the model and subsequent evaluation of its results: 87 600 TEC maps (80%) were allocated purposefully for the training stage, while the remaining 21 900 maps (20%) were used for the validation procedure the algorithm.

It should be emphasized that the vectors obtained on the hidden layer of the autoencoder may demonstrate the presence of relationships (correlate), which, in turn, may negatively affect the quality and interpretability of the decomposition. In this paper, to achieve orthogonality of the vectors, an additional structure was applied that initializes the weights of the hidden layer of the linear transformation using

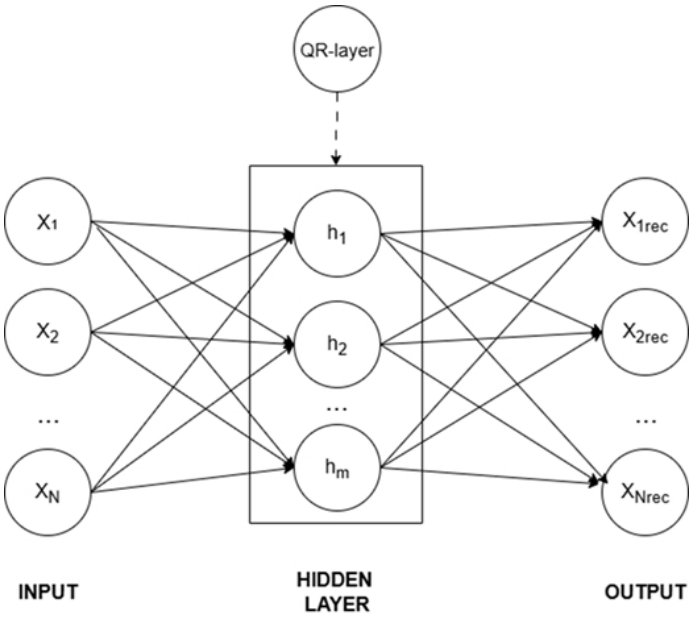

Fig. 1. The chosen architecture for compression of the original input feature space [9]

an orthogonal matrix obtained through the QR decomposition [9].

The number of hidden layers was 10, as it was shown earlier, they carry more than 95% of the initial information about the source data [8].

After completing the training, the encoder and decoder can be used independently. To form the feature space for subsequent machine learning models (for example, CatBoost), only a trained encoder was used. All the initial data (the time series of TEC and foF2) were passed through the encoder, and the corresponding 10-dimensional latent vectors were obtained at the output. These vectors, which are a compressed and informative representation of the initial state of the ionosphere, were later used as input features for forecasting algorithms. This approach allows not only to drastically reduce the dimensionality of the data, but also to transfer already refined and semantically saturated features to machine learning models, which ultimately increases the accuracy and reliability of the forecast.

Key technical parameters: the source data is compressed to 10 hidden vectors, the ReLU (Rectified Linear Unit) activation functions are used in the inner layers, batch normalization after each linear layer to stabilize learning and dropout (0.2) for regularization. The loss function consists of MSE (Mean Squared Error) for reconstruction. Optimization is performed using the AdamW method with L2 regularization (attenuation of the weights is 1e-5) and a learning rate of 0.001.

2. The database

2.1. Total electronic content data (TEC). Total Electron Content (TEC) maps provided by NASA's Jet Propulsion Laboratory (JPL) [10] (Fig. 2) were used as initial data on the state of the ionosphere. JPL GIM (Global Ionosphere Maps) data is one of the most accurate and widely used products in the global scientific community that characterize the integrated electronic content in the vertical column of the ionosphere. The maps are presented in IONEX format with a time resolution of 2 hours and a spatial resolution of 2.5° in longitude and 5° in latitude.

The preprocessing of the TEC time series for a particular location included the following steps.

1. Time series extraction: for a selected point with specified geographical coordinates, a time series of vertical TEC values was extracted from a sequence of global maps.
2. Identification and interpolation of gaps: the time series was checked for data gaps that could have been caused by failures in the satellite constellation or during processing. Minor gaps were restored using linear interpolation.

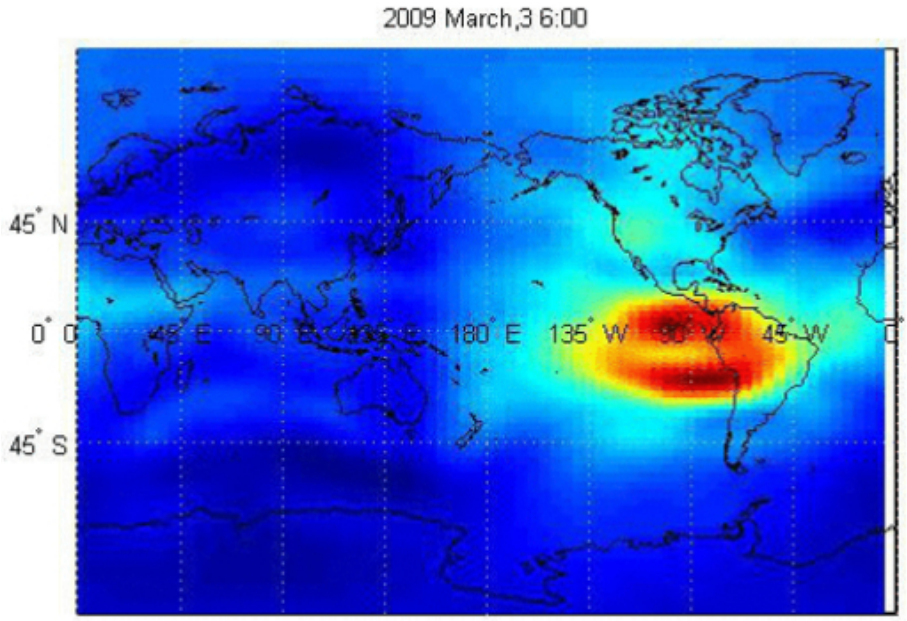

Fig. 2. JPL ionospheric TEC map [10] (color online)

3. Smoothing: to suppress high-frequency noise unrelated to large-scale ionospheric processes, a moving average with a window corresponding to 6 hours was used.
4. Sampling: the processed TEC time series served as the basis for constructing features describing the current state of the ionosphere.

2.2. Data on the values of the critical frequency of the F2 layer (foF2). To supplement the feature space with parameters directly characterizing the maximum electron density of the ionosphere, data on the critical frequency of the F2 layer (foF2) were used. These data were obtained from the global network of GIRO (Global Ionospheric Radio Observatory) ionosondes [11] (Fig. 3). GIRO provides access to the results of ionosphere vertical sounding from a large number of stations around the world, which allows for the acquisition of representative data on key ionosphere parameters, including foF2.

The preprocessing of foF2 data included the following steps.

1. Location and time matching: foF2 data was selected for the ionosonde station closest to the point for which the TEC series was extracted. The time stamps of the foF2 data were matched with the time stamps of the JPL TEC data.
2. Gap recovery: Since ionosonde data can have significant gaps (for example, due to technical malfunctions or poor radio propagation conditions), an interpolation algorithm for the nearest neighboring stations of the GIRO network was used to reconstruct the gap values, followed by time interpolation.

2.3. Formation of the final dataset. The processed TEC and foF2 time series were combined with data on solar (F10.7 index) and geomagnetic activity (Kp indices) obtained from open sources (DRAO, Leibniz Institute for Astrophysics Potsdam) [12, 13]. In earlier works, it was shown that they correlate with the extracted TEC series [14]. Cyclic time characteristics, such as time of year and time of day, were also added as features. The final dataset consisted of a table where each instance contained latent representations of TEC (enriched with foF2 values), external indices, and time stamps, allowing the use of classical machine learning algorithms for prediction.

Fig. 3. Location of the GIRO (Global Ionospheric Radio Observatory) network of global ionospheric stations (January 2020) [11]

3. Discussion of the results

Below is a flowchart of the proposed approach, which shows the main stages of data processing and modification (Fig. 4).

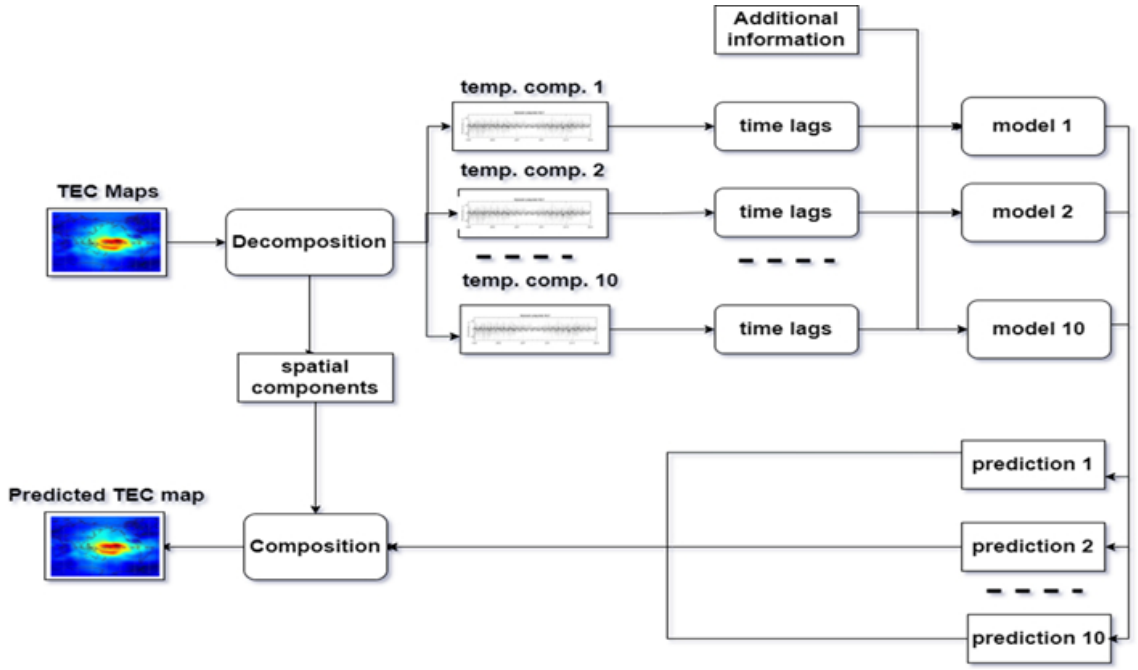

Fig. 4. Flowchart of the proposed approach [15]

3.1. Feature selection. In our case, the following features were taken into account, which can affect the prediction of the values of the time components of the TEC:

- time lags of the original series (over the past 72 hours), which allows us to capture changes in the electronic concentration (including fluctuations in values);
- the month, day, and hour of the date for which the forecast is made;
- the value of the F10.7 index at the time of the forecast;
- the value of the Kp index at the time of the forecast.

The time lags, which are listed first in the feature list, are essentially a hyperparameter that is adjusted during inference. It is important to find an optimal combination so that the number of lags does not significantly increase the prediction time, but can still improve accuracy. The time series of geophysical instrument readings have their own specificities, which should also be taken into account when selecting features for machine learning models. Table 1 below shows the SMAPE error and prediction time for the best algorithm (based on the results in Table 2).

The SMAPE metric is defined as

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i| + \varepsilon)/2}, \tag{1}$$

where y_i is the actual value, and $\hat{y}_i$ is the predicted value, $\varepsilon = 10^{-8}$ is used to ensure numerical stability for values close to zero (ensuring the universality of this expression).

SMAPE is a standard and widely accepted metric in time series forecasting, particularly in business analytics (predicting demand, sales, and load) [16].

3.2. Model training and validation. A critical aspect when working with time series is the correct separation of data into training and test samples. Unlike tasks where the data is independent

Table 1. Number of lags for the implementation of the forecasting algorithm (for the first component of the temporal decomposition of the autoencoder)

Number of lags, hours	SMAPE	Prediction time, s
24	0.21	0.35
48	0.14	0.72
72	0.06	0.91
96	0.06	1.2

and distributed equally, there is a strong time dependence (autocorrelation) in the time series. Arbitrary mixing and random division of data in this case would lead to a «leak of information from the future to the past»: the model, training from data from the entire time interval, could indirectly receive information about the future states of the system when predicting the past, which would distort the real assessment of its generalizing ability.

To avoid this, a strict temporary split was applied. The first 80% of the data (87600 vectors) corresponding to the initial time period were allocated to the training sample. The remaining 20% of the data (21900 vectors), chronologically following the training sample, made up the test sample. This approach ensures that the model is trained solely on past data and tested on future data, which fully simulates the conditions of real operation and ensures adequate validation of its predictive effectiveness.

To compare the models at the training and validation stage, the following metrics were used: MAE (mean absolute error), MSE (mean squared error), RMSE (root mean squared error), coefficient of determination R^2 and the SMAPE already described above.

The algorithms were compared as the average error of all 10 models (for each of the components) for the forecast for 24 hours ahead.

Table 2. Machine learning model training results (top 5 best algorithms)

Algorithm	MAE	MSE	RMSE	R^2	SMAPE
CatBoost	1.9	11.8	3.4	0.98	0.25
XGBoost	2.1	15.2	3.9	0.98	0.27
LightGBM	3.1	17.2	4.1	0.98	0.28
ExtraTrees	3.2	17.2	4.1	0.97	0.29
Random Forest	3.4	18.2	4.2	0.97	0.30

Table 2 shows that one of the modifications of the CatBoost gradient boosting algorithm showed the best results for the features obtained using the neural network of the autoencoder.

Previously, the authors developed a dimensionality reduction model for TEC data based on the linear principal component analysis (PCA), which has shown its basic effectiveness relative to the models used [15]. However, the linear nature of the transformation in PCA may limit its ability to identify complex nonlinear patterns inherent in ionospheric dynamics. To quantify the advantages of the new approach, a comparative analysis of the predictive effectiveness of hybrid models using both the proposed autoencoder and the previously developed PCA-based model for feature generation was carried out (the number of compressed vectors in both cases was 10).

To assess the quality of the restored maps, the WMAPE metric was used, which allows us to take

into account the spatial contribution of each individual predicted point on the map:

$$\mathrm{WMAPE} = \frac{\sum_{i=1}^{n} |y_i| \frac{|y_i - \hat{y}_i|}{|y_i|}}{\sum_{i=1}^{n} |y_i|}. \tag{2}$$

The comparison results are shown in Table 3, where two approaches are compared: based on the principal component analysis (PCA) and the autoencoder (AE) [15].

The WMAPE values presented in Table 3 are calculated for the global coverage area of the TEC maps. The averaging is performed over all spatial cells of the calculation grid. This approach to error aggregation, unlike point or regional estimation, allows for the consideration of its spatial heterogeneity across the entire study area. This is particularly critical for analyzing the quality of models in key regions of the ionosphere, such as the equatorial anomaly zone, where prediction errors often exhibit systematic and elevated patterns. Consequently, the global WMAPE provides an integrated and most representative assessment of model performance for the task of field reconstruction.

Table 3. Results of comparison of reconstructed maps based on two approaches

Forecast period, hours	МГК	АК
2	0.21	0.15
4	0.25	0.17
12	0.27	0.20
24	0.31	0.23
48	0.41	0.31
72	0.51	0.35
144	0.61	0.42

Conclusions and summary

In the course of the conducted research, a two-stage hybrid model was successfully developed and tested for predicting ionosphere parameters, in particular, total electron content (TEC). The key element of the approach was the use of a fully-connected autoencoder for nonlinear dimensionality reduction of the initial high-dimensional data, which combines time series of TEC maps and critical frequency values of the F2 layer (foF2). This allowed for the effective compression of information, the extraction of the most significant latent features, and the elimination of noise, resulting in a compact 10-dimensional space for describing the ionosphere's state.

In the second stage, it was demonstrated that the obtained latent representations, enriched with data on solar and geomagnetic activity, are a highly effective feature space for classical machine learning algorithms. A comparative analysis showed that the CatBoost gradient boosting algorithm demonstrates the best prediction accuracy on the test set, as confirmed by the RMSE and MAE metrics. An important condition for a reliable assessment was the correct data time separation procedure, which eliminated information leakage.

Thus, the proposed hybrid architecture overcomes the limitations of both purely physical and empirical models, as demonstrated by the WMAPE metric values. This may indicate its potential superiority compared to classical methods. The combination of deep neural networks for feature extraction and ensemble methods for regression opens up a promising direction for creating accurate and operational ionosphere state prediction systems, which is crucial for ensuring the reliability of satellite navigation and communication.

References

1. Brjunelli BE, Namgaladze AA. Physics of the Ionosphere. M.: Nauka; 1988. 527 p. (in Russian).
2. Mendillo M. Storms in the ionosphere: Patterns and processes for total electron content. Rev. Geophys. 2006;44(4):RG4001. DOI: 10.1029/2005RG000193.

3. Maksimov DS, Kogogin DA, Nasyrov IA, Zagretdinov RV. Solar flares in the 25th cycle of activity: Effect on ionospheric disturbance and GNSS signal strength. Current Problems of Remote Sensing of the Earth from Space. 2025;22(3):301–317 (in Russian). DOI: 10.21046/ 2070-7401-2025-22-3-301-317.
4. Fitzgerald TJ. Observations of total electron content perturbation of GPS signals caused by a ground level explosion. Journal of Atmospheric and Solar-Terrestrial Physics. 1997;59(7): 829–834. DOI: 10.1016/s1364-6826(96)00105-8.
5. Feng JD, Zhang T, Li W, Zhao Zh, Han B, Wang K. A new global TEC empirical model based on fusing multi-source data. GPS Solutions. 2023;27(1):20. DOI: 10.1007/s10291-022-01355-8.
6. Bilitza D., Pezzopane M., Truhlik V., Altadill D., Reinisch B. W., Pignalberi A. The international reference ionosphere model: A review and description of an ionospheric benchmark. Rev. Geophys. 2022;60(4):e2022RG000792. DOI: 10.1029/2022RG000792.
7. Natras R, Soja B, Schmidt M. Ensemble machine learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content forecasting. Remote Sens. 2022;14(15):3547. DOI: 10.3390/rs14153547.
8. Appalonov AM, Maslennokova YuS, Sherstyukov ON. Application of deep learning neural networks for the analysis of spatial and temporal components of the decomposition of the total elec-tronic content of the ionosphere. Radioengineering. 2025;89(1):172–179 (in Russian). DOI: 10.18127/j00338486-202501-16.
9. Appalonov AM, Maslennikova JuS, Sherstyukov ON. Analysis of spatiotemporal variations of the complete electronic content and the critical frequency of the F2 layer using deep learning neural networks. Propagation of radio waves [Electronic resource]. In: Collection of Reports of the XXIX All-Russian Open Scientific Conference. June 30–July 4, 2025, Kazan, Russia. Kazan: Kazan University Publishing, 2025. P. 512–515. Available from: https://repository.kpfu.ru/?p_id=319113.
10. JPL. Official site of JPL [Electronic resource]. Available from: https://www.jpl.nasa.gov.
11. Global Ionospheric Radio Observatory. Official site[Electronic resource]. Available from: https://giro.uml.edu/.
12. Dominion Radio Astrophysical Observatory (DRAO) : official site / National Research Council Canada [Electronic resource]. Available from: https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca.
13. Helmholtz-Zentrum Potsdam – Deutsches GeoForschungsZentrum (GFZ). Official site / Helmholtz Association [Electronic resource]. Available from: https://www.gfz.de/.
14. Appalonov AM, Maslennikova JuS, Sherstyukov ON. Decomposition of global maps of the complete electronic content of the ionosphere using neural networks. In: Proceedings of the 22nd International Conference “Modern Problems of Remote Sensing of the Earth from Space”. M.: Russian Space Research Institute Publishing; 2024. P. 434 (in Russian). DOI: 10.21046/22DZZconf-2024a.
15. Appalonov AM, Maslennikova YuS. Short-term prediction of the total electronic content of the ionosphere using solar parameters by machine learning methods. In: Proceedings of the 21 International Conference “Modern Problems of Remote Sensing of the Earth from Space”. M.: Russian Space Research Institute Publishing; 2023. P. 299 (in Russian). DOI: 10.21046/ 21DZZconf-2023a.
16. Masini RP, Medeiros MC, Mendes EF. Machine learning advances for time series forecasting. Journal of Economic Surveys. 2021;37(1):76–111. DOI: 10.1111/joes.12429.