engineering proceedings

MDPI

Proceeding Paper

Natural Data Analysis Method Based on Wavelet Filtering and NARX Neural Networks †

Oksana Mandrikova, Yurii Polozov and Bogdana Mandrikova *

Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS, 684034 Paratunka, Kamchatskiy Krai, Russia; oksanam1@mail.ru (O.M.); polozov@ikir.ru (Y.P.)
* Correspondence: 555bs5@mail.ru; Tel.: +7-41531-33193; Fax: +7-41531-33718
† Presented at the 15th International Conference "Intelligent Systems" (INTELS'22), Moscow, Russia, 14–16 December 2022.

Abstract: A method for analyzing natural data and detecting anomalies is proposed. The method is based on combining wavelet filtering operations with the NARX neural network. The analysis of natural data and the detection of anomalies are of particular relevance in the problems of geophysical monitoring. An important requirement of these methods is their adaptability, accuracy and efficiency. Efficiency makes it possible to detect anomalies timely in order to prevent catastrophic natural phenomena. Wavelet filtering operations include the application of a multi-scale analysis construction and threshold functions. The article proposes a wavelet filtering algorithm and a method for estimating thresholds based on a stochastic approach. The operations of the method implementation are described. It is shown that the use of wavelet filtering allows one to suppress noise, simplifies the data structure and, as a result, allows one to obtain a more accurate NARX neural network model. The effectiveness of the method for detecting ionospheric anomalies during periods of magnetic storms is shown using the data of the critical frequency of the ionosphere as an example.

Keywords: data analysis; wavelets; neural networks; ionosphere

check for updates

Citation: Mandrikova, O.; Polozov, Y.; Mandrikova, B. Natural Data Analysis Method Based on Wavelet Filtering and NARX Neural Networks. *Eng. Proc.* **2023**, *33*, 63. https://doi.org/10.3390/engproc2023033063

Academic Editors: Askhat Diveev, Ivan Zelinka, Arutun Avetisyan and Alexander Ilin

Published: 16 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Investigation of natural data is a meticulous, nontrivial task and is of great concern in the research of different processes and phenomena. Natural data processing and analysis are widely applied in different fields such as technical processes, geophysics, biology, economy, chemistry and so on. In the context of this problem, methods of object statistical analysis and anomaly detection are of special interest. Examples might include the problems of detection of outer space disturbances [1], floods, earthquakes [2], mud flows and landslides, tsunamis [3], changes in the geological environment [4] and other natural emergency situations. The problem of anomaly recognition is also very important in medicine. Timely and highly accurate diagnostics make it possible to detect and identify patient physical state disorders that in some cases may cost someone their life. Such methods should be fast and flexible in order to detect rapidly changing states of systems and objects. These changes characterize anomaly occurrences.

Different methods are used for natural data analysis. A wide spectrum of approaches to this problem is determined by the complexity of the data under investigation and includes deterministic [5] and stochastic [6] tools. At the present time, different combinations of such approaches are being actively developed [7]. As methods for natural data analysis developed, it became evident that classical approaches to modeling and analysis (AR, ARMA models [8], exponential smoothing [9], stochastic approximation [6], etc.) do not allow us to describe properly the complicated nonstationary structure of natural time series.

At the present time, hybrid approaches and the methods [7] allowing one to improve data analysis quality are developing rapidly to solve such problems. Wavelet transform is a flexible instrument and is widely used in the problems of data processing and analysis.

A large library of wavelet filters and a wide choice of constructions for signal expansion provide the possibility to adapt this instrument for the data of different structures. For example, we developed the empirical wavelet filters [10], which are effective in image processing applications.

Now, neural network methods [11] are of great importance. They are known to be able to approximate complex non-linear dependencies in data and automate processes. However, to achieve high efficiency of the neural network apparatus, it is necessary to have high-quality training data and their representativeness. The article proposes a method based on combining the operations of wavelet filtering and NARX neural networks. NARX networks are an evolution of the ARIMA class of models. They approximate non-linear time series and may have different architectures. This allows you to flexibly adapt to the tasks of data analysis and forecasting. Wavelet filtering is used to reduce noise and simplify the data structure. This improves the efficiency of the neural network. Similar hybrid approaches to the analysis of complex data have already been used in other works [7,12]. One such work is the combination of ARIMA and discrete wavelet decomposition together with the LSTM neural network [7]. Another example is the work [12], which proposes a discrete wavelet decomposition with ARIMA and a neural network. Based on this method [12], the authors carry out a forecast of hydrological hazards (high water levels (floods), floods, rain floods, etc.). In this paper, the approach of combined application of wavelet filtering and NARX networks is used to improve the efficiency of autoregressive time series models. With the use of wavelet filtering, it is possible to reduce noise and simplify the structure of the initial data. The article shows that this leads to an increase in the performance of the neural network. The wavelet filtering process includes a combination of multiscale analysis [10] and threshold functions.

The article presents the wavelet filtering algorithm and the threshold estimation method. Thresholds were estimated based on statistical operations. The study of the time series of the critical frequency of the ionospheric layer F2 (foF2) was carried out. The structure of the recorded ionospheric data depends on many temporal, weather, positional, and other factors [13]. Time series of the ionosphere demonstrate a regular course and various anomalies in the form and duration. Anomalies can be observed under conditions of increased solar and geomagnetic activity, as well as during seismic activity [13].

It is known that there are cases of ionospheric anomaly negative impacts on high-frequency radio communication and navigation signals from GPS satellite. One example is an event in January 2005. Due to space weather anomalies, 26 flights of United Airlines were redirected to non-optimal routes in order to avoid catastrophic risks of communication losses. An event in October 2003 is also known when the American Wide Area Augmentation System remained inoperative for more than one day and did not navigate vertically because of ionospehric inhomogeneities [14]. As it was mentioned above, traditional approaches to the analyzed ionospheric parameters do not satisfy the requirements of anomaly detection accuracy. In this paper, we suggest a method that confirmed its efficiency in the task of detection of anomalies (ionospehric inhomogeneities) generated by magnetic storms. It was empirically proved that signal pre-processing based on wavelet filtering improves the accuracy of NARX neural network model for ionospheric parameter time series. The paper presents a comparison of the efficiency of the proposed method with direct application of NARX neural network.

2. Applied Data

In this paper, we applied the ionospheric critical foF2 data [13], which have a one-hour sampling rage. The data have been recorded at the Paratunka site (Kamchatskiy kray, Russia, site coordinates: 53.0 N, 158.7 E) since 1969. Due to instrumentation failures and natural and man-made factors, there are gaps in foF2 data. We used a time series with data gaps no longer than one day. When there were gaps in the data of less than 25 h, they were filled in based on the median method.

3. Method Description

3.1. Application of Wavelet Transform and Threshold Function

Natural time series contain noise formed under a large variety of natural and man-made factors [15]. To improve the quality of natural data analysis based on neural networks, according to the papers [16], we applied noise suppression operations. They included the application of a multiscale analysis construction [17] and threshold function. The algorithm for noise suppression is the following:

1. Signal $f_0(t)$ wavelet decomposition into the components:

$$f_0(t) = \sum_{j=-1}^{-m} g_j(t) + f_{-m}(t),$$

where $f_{-m}(t) = \sum_k c_{-m,k}\phi_{-m,k}(t)$ is the smoothed component, $c_{-m,k} = \langle f_0, \phi_{-m,k}\rangle$, $\phi_{-m,k}(t) = 2^{-m/2}\phi(2^{-m}t - k)$ is the scaling function, $g_j(t) = \sum_k d_{j,k}\Psi_{j,k}(t)$ are detailing components, $d_{j,k} = \langle f_0, \Psi_{j,k}\rangle$, $\Psi_{j,k}(t) = 2^{j/2}\Psi(2^j t - k)$ is the wavelet, j is the decomposition level, the decomposition level $j = 0$ is assumed for the initial signal.

2. Application of the threshold fucntion to the component coefficients $g_j(t)$:

$$T(d_{j,k}) = \begin{cases} 0, \ if \ |d_{j,k}| \le T_j \\ d_{j,k}, \ if \ |d_{j,k}| > T_j \end{cases}, \quad (1)$$

where $T_j = t_{1-\frac{\alpha}{2},N-1}\hat{\sigma}_j$, $t_{\alpha,N}$ are α-quantiles of Student's distribution, $\hat{\sigma}_j$ is the sample standard deviation, the decomposition levels $j = \overline{-1,-m}$.

3. Signal wavelet recovery:

$$\widetilde{f}_0(t) = \sum_{j,k} T(d_{j,k})\Psi_{j,k}(t) + f_{-m}(t).$$

Determination of thresholds T_j in operation (1) was based on the paper results [17] and it was assumed that absolute values of the coefficients $|d_{j,k}|$ of the components $g_j(t)$ were close to zero out of the neighborhood containing signal structural characteristics. In this case, the probability ($\alpha \approx 0.99$) that the coefficients $|d_{j,k}|$ with respect to the argument k are within the interval $(\mu - 3\sigma; \mu + 3\sigma)$ is high. Here, $\mu \approx 0$ is the mathematical expectation of the value $|d_{j,k}|$, σ is the standard deviation (three-sigma rule [18]). Then, for normally distributed $|d_{j,k}|$, we can estimate the thresholds T_j with defined confidence level α as $T_j = t_{1-\frac{\alpha}{2},N-1}\hat{\sigma}_j$, where $t_{\alpha,N}$ are the α-quantiles of the Student's distribution [18].

3.2. Application of NARX Neural Networks

After noise suppression operations, data are analyzed based on NARX neural networks [11]. We applied feedback NARX networks [11], the architectures of which are illustrated in Figure 1.

The value vector applied to the hidden layer neurons consists of the following components:

- Current $\widetilde{f}_0(t)$ and previous $\widetilde{f}_0(t-1), ..., \widetilde{f}_0(t-l_x)$ values of the input vector;
- Output values during the previous time intervals $\hat{f}_0(t), \hat{f}_0(t-1), ..., \hat{f}_0(t-l_y)$.

The neural network input value $\hat{f}_0(t+1)$ has the form:

$$\hat{f}_0(t+1) = F[\widetilde{f}_0(t), \widetilde{f}_0(t-1), ..., \widetilde{f}_0(t-l_x), \hat{f}_0(t), \hat{f}_0(t-1), ..., \hat{f}_0(t-l_y)], \quad (2)$$

where $F(\cdot)$ is the neural network display function.

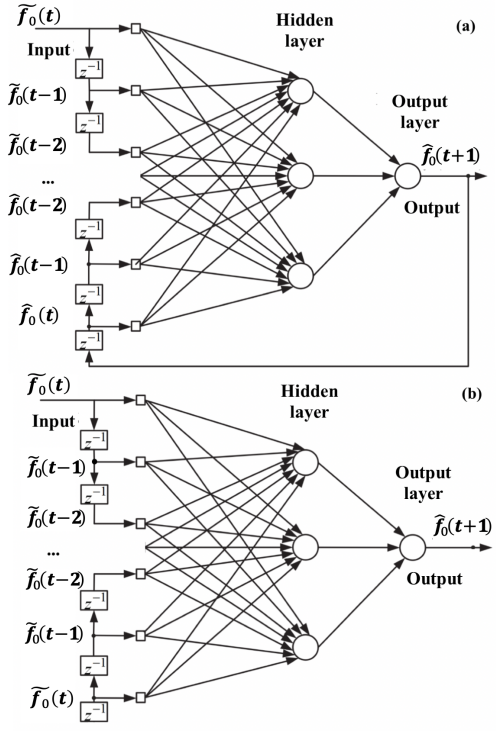


Figure 1. Architecture of a generalized recurrent network. (**a**) NARX PA architecture; (**b**) NARX SPA architecture.

Figure 1 shows the NARX network with Series-Parallel Architecture (NARX SPA) (Figure 1b) and NARX network with Parallel Architecture (NARX PA) (Figure 1a) [11]. Analytical form for the NARX PA architecture is represented in expression (2). In NARX SPA, previous values $\widetilde{f}_0(i)$ are applied to the input instead of the outputs $\hat{f}_0(i), i = \overline{t, t-l_y}$. In this case, the network output value $\hat{f}_0(t+1)$ has the form

$$\hat{f}_0(t+1) = F[\widetilde{f}_0(t), \widetilde{f}_0(t-1), ..., \widetilde{f}_0(t-l_x), \widetilde{f}_0(t), \widetilde{f}_0(t-1), ..., \widetilde{f}_0(t-l_y)].$$

It is known from [11], that NARX SPA are applied when data of a previous value vector are available. NARX PA networks are applied when only modeling results can be used, for example, when making a forecast with a large step ahead.

The network memory parameters were used in the experiments: $l_x = l_y = 2, l_x = l_y = 5$.

The Bayesian Regularization algorithm with Levenberg–Marquardt optimization was used to train the networks [11]. That allows us to minimize the linear combination of error squares and weights. After that, their appropriate combination is formed in such a way that the network has the highest degree of generalization.

3.3. Scheme of Method Realization

The scheme of method realization is illustrated in Figure 2. Anomalies are detected based on estimation of neural network errors

$$\varepsilon_i = \sum_{i=i-s}^{i+s} |\hat{f}_0(i) - \widetilde{f}_0(i)|. \quad (3)$$

There is an anomaly in the data if $\varepsilon_i > 2\sigma + \varepsilon_{mean}$, where σ is standard deviation of errors ε_i during periods without anomalies, and ε_{mean} is mean of errors ε_i in periods without anomalies.

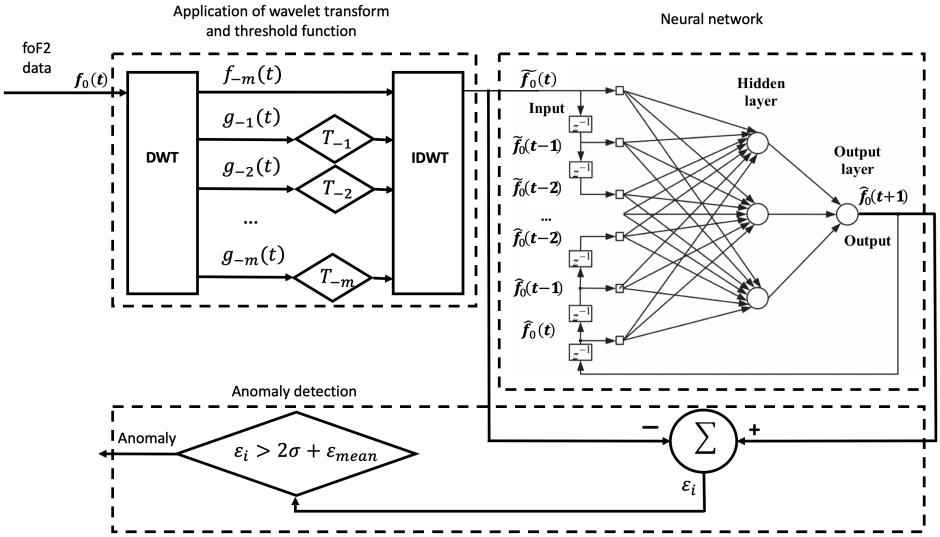


Figure 2. Scheme of method realization.

4. Results of Method Application for Ionospheric Data

Regular foF2 data depend on the time of year and the level of solar activity. Therefore, separate training samples were formed (to build a neural network) for different seasons and levels of solar activity. Two periods of solar activity were studied in the work: high and low. The level of solar activity was estimated based on the average monthly values of the solar index f10.7. We used the ionospheric data recorded during a calm geomagnetic state for the period 1969–2015. When constructing each neural network, training sampling contained about 2000 counts. According to the suggested method, neural networks were trained using the data obtained after wavelet filtering. When making wavelet filtering operations, orthonormal Daubeches third-order wavelets were applied [10]. They were determined by minimizing the data approximation errors. To estimate the method, neural networks were also trained via applying the foF2 initial data. We constructed 16 neural networks using the NARX PA and NARX SPA architectures, which had the memory $l_x = l_y = 2, l_x = l_y = 5$.

Tables 1 and 2 show the results of the obtained neural network operation. Root-mean-square (RMS) deviations of network errors were determined as

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(e_i - \bar{e})^2},$$

where $\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i$, $e_i = \hat{f}_0(i) - \widetilde{f}_0(i)$.

Table 1. Standard deviations of NARX neural network errors with delays (2) and (5). T_j is the wavelet filtering.

Season	SPA (2)	PA (2)	SPA (2) (T_j)	PA (2) (T_j)	SPA (5)	PA (5)	SPA (5) (T_j)	PA (5) (T_j)
Winter (low solar activity)	0.47	0.57	0.25	0.41	0.42	0.43	0.18	0.30
Summer (low solar activity)	0.42	0.46	0.19	0.29	0.36	0.39	0.16	0.26
Winter (high solar activity)	0.50	0.78	0.32	0.51	0.43	0.61	0.28	0.47
Summer (high solar activity)	0.42	0.45	0.26	0.29	0.36	0.37	0.21	0.22

Table 2. Results of operation of neural networks (Summer, low solar activity). T_j is the wavelet filtering.

Network Memory $l_x = l_y$	Network Architecture	ε_{mean} (T_j)	σ (T_j)	ε_{mean}	σ
2	NARX SPA	0.43	0.37	0.98	0.48
2	NARX PA	0.85	0.48	1.3	0.6
5	NARX SPA	0.33	0.28	0.83	0.37
5	NARX PA	0.80	0.42	1.27	0.56

An analysis of the results from Table 1 shows that NARX SPA architecture describes more adequately the foF2 data time series. When using a large number of delay lines, the results become more comparable in favor of NARX SP (Table 1). A comparison of the results in Table 2 shows that application of wavelet filtering operation allows us to significantly increase the neural network operation quality that confirms the suggested method efficiency.

To test the adequacy of neural network models, the autocorrelation of network errors was evaluated based on the Ljung-Box test $Q = M(M+2)\sum_{s=1}^{L}\frac{p_s^2}{T-s}$, where M is the observation number, p_s is the autocorrelation of the s-th order, and L is the checked lag number. If $Q > \chi^2_{1-\alpha,L}$, where $\chi^2_{1-\alpha,L}$ are the distribution quantiles chi-square with L freedom degree, the presence of autocorrelation of the L-th order in the time series is recognized.

An analysis of the results in Table 3 shows that Ljung–Box test values for the networks constructed without wavelet filtering significantly exceed the critical values $\chi^2_{1-\alpha,L}$. The results indicate a significant correlation of errors of these neural networks and, as a consequence, insufficient quality of foF2 data approximation. When the network memory is increased, data approximation quality improves. The networks, trained applying data wavelet filtering operations, show the best results. For a small memory of networks $l_x = l_y = 2$, in accordance with the Ljung–Box test, network errors are uncorrelated. This confirms the adequacy of the obtained neural network model.

Table 3. Estimate of error autocorrelation of neural networks (Summer, low solar activity). T_j is the wavelet filtering.

Network Memory $l_x = l_y$	Lag Number L	Q (T_j)	Q	$\chi^2_{1-\alpha,L}$
2	1	4.19	3.82	3.84
2	6	13.53	48.18	12.59
2	12	41.41	98.97	21.02
5	1	1.17	0.51	3.84
5	6	11.17	15.53	12.59
5	12	26.38	44.6	21.02

Figure 3 shows the results of neural network operation during a magnetic storm, which occurred on 16 July 2017. The red dashed line in Figure 3 marks the magnetic storm beginning. According to Space Weather site data (http://ipg.geospace.ru, accessed on 1 March 2023), the magnetic storm had a mixed nature, arrival of an accelerated flux from a coronal hole and coronal mass ejection. During strong geomagnetic disturbances, the DST index (DST index of geomagnetic activity) reached the minimum of -72 nTl (Figure 3f,l). Analysis of foF2 initial data (Figure 3a,g) and comparison with the moving median (in Figure 3a,g, the median is a green dashed line) show the distortion of data regular time variation on July 16. That is determined by ionospehric disturbance occurrences. Anomalous changes in foF2 data on July 16 confirm the results of the neural networks, as a significant error increase is observed (Figure 3b–e,h–k). However, as the analysis shows, the application of wavelet filtering improves the network performance quality. The errors of the network with small memory $l_x = l_y = 2$ are close to zero except for the anomalous period.

During the anomalous period, network errors significantly increase and the anomaly is clearly detected.

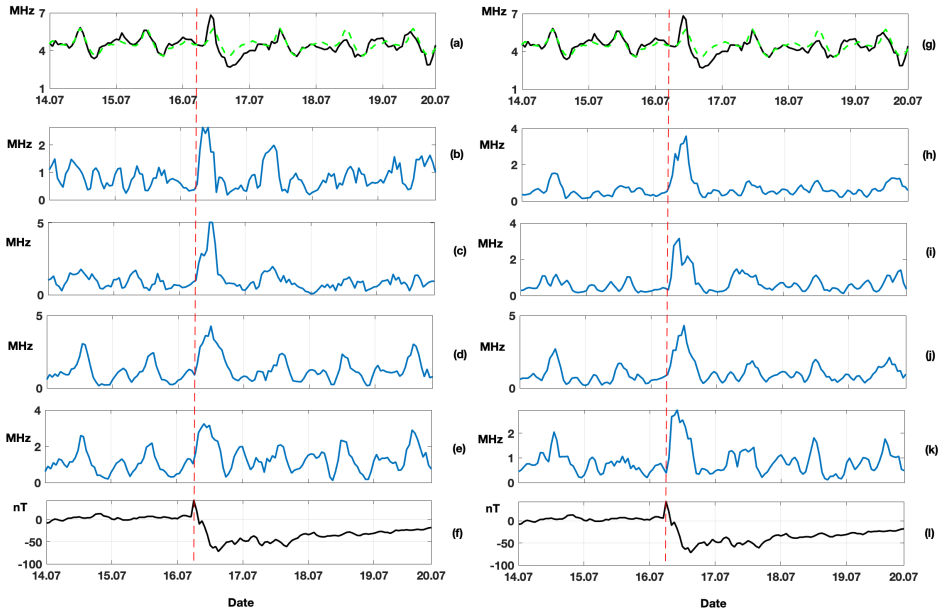


Figure 3. Result of data processing foF2. (**a**,**g**) foF2 data (black) and foF2 median (green); (**b**,**c**) NARX SPA errors with memory 2 and 5, respectively; (**d**,**e**) NARX PA errors with memory 2 and 5, respectively; (**h**,**i**) NARX SPA errors with wavelet filtering and memory 2 and 5, respectively; (**j**,**k**) NARX PA errors with wavelet filtering and memory 2 and 5, respectively; (**f**,**l**) DST. Red dashed line is the magnetic storm beginning.

When comparing the results of NARX PA and NARX SPA architectures, we find that for the delay lines $l_x = l_y = 2, l_x = l_y = 5$, NARX SPA architecture detects anomalous changes more clearly both when applying initial data to the network input and when applying the data after wavelet filtering to the input. A comparison of the results of NARX SPA neural networks with moving median confirms the effectiveness of the method. The moving median has errors in the period after the storm on July 18, 2017, which are absent in the neural network model.

5. Conclusions

The application of wavelet filtering together with a NARX neural network is effective for the task of detection of ionospehric anomalies. Direct application of NARX neural networks does not allow one to approximate ionospheric time variation due to the presence of long-term dependencies. Wavelet filtering operation and application of stochastic thresholds significantly decrease the error level even for small memory networks. Wavelet filtering suppresses noise, simplifies the data structure and, as a consequence, makes it possible to obtain a more accurate NARX neural network model. The application of the Ljung–Box test showed no correlation of network errors and confirmed the adequacy of the obtained neural network model.

Comparison of NARX SPA and NARX PA showed the advantage of the NARX SPA architecture, which makes it possible to obtain a more adequate model for describing the ionospheric time series. Using the example of a strong magnetic storm, the possibility of the method for detecting ionospheric anomalies is shown. Comparison of the NARX SPA network with the moving median, widely used to analyze ionospheric data, showed the efficiency of the proposed approach.

With an extension of the statistics and a more detailed study of ionospheric data during anomalous periods, the authors plan to continue research in this direction.

Author Contributions: Conceptualization, O.M. and Y.P.; methodology, O.M., Y.P. and B.M.; software, Y.P.; validation, O.M. and Y.P.; formal analysis, O.M. and Y.P.; investigation, O.M. and Y.P.; resources, Y.P.; data curation, Y.P.; writing—original draft preparation, Y.P. and B.M.; writing—review and editing, O.M.; visualization, O.M. and Y.P.; project administration, O.M. All authors have read and agreed to the published version of the manuscript.

Funding: The work was carried out as a part of the implementation of the State Task AAAA-A21-121011290003-0. The work was carried out by means of the Common Use Center "North-Eastern Heliogeophysical Center".

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tang, R.; Zeng, F.; Chen, Z.; Wang, J.-S.; Huang, C.-M.; Wu, Z. The comparison of predicting storm-time ionospheric TEC by three methods: ARIMA, LSTM, and Seq2Seq. *Atmosphere* **2020**, *11*, 316. [CrossRef]
2. Perol, T.; Gharbi, M.; Denolle, M. Convolutional neural network for earthquake detection and location. *Sci. Adv.* **2018**, *4*, e1700578. [CrossRef] [PubMed]
3. Kim, S.-K.; Lee, E.; Park, J.; Shin, S. Feasibility analysis of GNSS-reflectometry for monitoring coastal hazards. *Remote Sens.* **2021**, *13*, 976. [CrossRef]
4. Alperovich, L.; Eppelbaum, L.; Zheludev, V.; Dumoulin, J.; Soldovieri, F.; Proto, M.; Bavusi, M.; Loperte, A. A new combined wavelet methodology: Implementation to GPR and ERT data obtained in the Montagnole experiment. *J. Geophys. Eng.* **2013**, *10*, 025017. [CrossRef]
5. Chen, J.; Heincke, B.; Jegen, M.; Moorkamp, M. Using empirical mode decomposition to process marine magnetotelluric data: Using EMD to process marine MT data. *Geophys. J. Int.* **2012**, *190*, 293–309. [CrossRef]
6. Robbins, H.; Monro, S. A stochastic approximation method. *Ann. Math. Stat.* **1951**, *22*, 400–407. [CrossRef]
7. Wu, X.; Zhou, J.; Yu, H.; Liu, D.; Xie, K.; Chen, Y.; Hu, J.; Sun, H.; Xing, F. The development of a hybrid Wavelet-ARIMA-LSTM model for precipitation amounts and drought analysis. *Atmosphere* **2021**, *12*, 74. [CrossRef]
8. Box, G.E.P.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*, Rev. ed.; Holden-Day series in time series analysis and digital processing; Holden-Day: San Francisco, CA, USA, 1976.
9. Liu, J.; Kumar, S.; Palomar, D.P. Parameter estimation of heavy-tailed AR model with missing data via stochastic EM. *IEEE Trans. Signal Process.* **2019**, *67*, 2159–2172. [CrossRef]
10. Mallat, S.G. *A Wavelet Tour of Signal Processing: The Sparse Way*; Elsevier: Amsterdam, The Netherlands, 2009.
11. Haykin, S.S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 1999.
12. Phan, T.-T.-H.; Nguyen, X.H. Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. *Adv. Water Resour.* **2020**, *142*, 103656. [CrossRef]
13. Danilov, A.D. Ionospheric F-region response to geomagnetic disturbances. *Adv. Space Res.* **2013**, *52*, 343–366. [CrossRef]
14. Kuznetsov, V.D. Space weather and risks of space activity. *Space Eng. Technol. Mag.* **2014**, *3*, 3–13
15. Geppener, V.V.; Mandrikova, B.S. Detecting and identifying anomalous effects in complex signals. *Autom. Remote Control* **2021**, *82*, 1668–1678. [CrossRef]
16. Mandrikova, O.; Mandrikova, B. Hybrid method for detecting anomalies in cosmic ray variations using neural networks autoencoder. *Symmetry* **2022**, *14*, 744. [CrossRef]
17. Chui, C.K. *An Introduction to Wavelets*; Academic Press: Boston, MA, USA, 1992
18. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed.; Springer series in statistics; Springer: New York, NY, USA, 1993.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.