




LOCAL DIAGNOSTICS OF AURORA PRESENCE BASED ON INTELLIGENT ANALYSIS OF GEOMAGNETIC DATA

A.V. Vorobev 
Geophysical Center RAS,
Moscow, Russia, geomagnet@list.ru
Ufa University of Science and Technology,
Ufa, Russia
Ufa State Petroleum Technological University,
Ufa, Russia

A.A. Soloviev 
Geophysical Center RAS,
Moscow, Russia, a.soloviev@gcras.ru
Schmidt Institute of Physics of the Earth RAS,
Moscow, Russia

V.A. Pilipenko 
Geophysical Center RAS,
Moscow, Russia, pilipenko_va@mail.ru
Schmidt Institute of Physics of the Earth RAS,
Moscow, Russia

G.R. Vorobeva 
Ufa University of Science and Technology,
Ufa, Russia, gulnara.vorobeva@gmail.com
Ufa State Petroleum Technological University,
Ufa, Russia

A.A. Gainetdinova
Ufa University of Science and Technology,
Ufa, Russia, gainetdinova.aa@ugatu.su

A.N. Lapin 
Ufa University of Science and Technology,
Ufa, Russia, meccos160@yandex.ru

V.B. Belakhovsky
Polar Geophysical Institute RAS,
Apatity, Russia, belakhovsky@mail.ru

A.V. Roldugin
Polar Geophysical Institute RAS,
Apatity, Russia, roldugin_a@pgia.ru

Abstract. Despite the existing variety of approaches to monitoring space weather and geophysical parameters in the auroral oval region, the issue of effective prediction and diagnostics of auroras as a special state of the upper ionosphere at high latitudes remains virtually unresolved.

In this paper, we explore the possibility of local diagnostics of auroras through mining of geomagnetic data from ground-based sources. We assess the significance of indicative variables and their statistical relationship.

So, for example, the application of Bayesian inference to the data from the Lovozero geophysical station for 2012–2020 has shown that the dependence of a posteriori probability of observing auroras in the optical range on the state of geomagnetic parameters is logarithmic, and the degree of its significance is inversely proportional to the discrepancy between empirical data and approximating function.

The accuracy of the approach to diagnostics of aurora presence based on the random forest method is at least 86 % when using several local predictors and ~80 % when using several global geomagnetic activity indices characterizing the geomagnetic field disturbance in the auroral zone.

In conclusion, we discuss promising ways to improve the quality metrics of diagnostic models and their scope.

Keywords: auroras, geomagnetic variations, geomagnetic data, ascaplots, machine learning, data mining, Bayesian inference, random forest.

INTRODUCTION

As is known, the highest risks in decreasing the level of technosphere safety associated with the effects of space weather on high-latitude infrastructure facilities (failures in HF radio communication systems and railway automation, additional errors in magnetic inclinometers, failures in power systems, reduction of life of main pipelines due to an increase in their corrosion rate, etc. [Sokolova et al., 2019; Ptitsyna et al., 2008; Vorobev et al., 2022a; Soloviev et al., 2022; Pilipenko, 2021]) are assessed in the auroral oval region — a belt of intense

aurora created by electron precipitation from near-Earth space into the atmosphere. It is in this region that, due to its characteristic sharp gradients and a high level of turbulence of ionospheric plasma, the most frequent navigation signal phase failures and extreme positioning errors are recorded [Zakharov et al., 2020]. As a result, the error in high-precision navigation of GPS receivers in PPP mode (Precise Point Positioning), operated in the region of auroral electron precipitation into the ionosphere, can increase up to five times relative to the background level [Yasyukevich et al., 2018, 2020].

There are periodic reports from NATO and the Russian Aerospace Defense Forces on global failures in GPS signal receiving systems employed in the auroral oval zone [<https://www.gpsworld.com/norway-finland-suspect-russia-of-jamming-gps>]. For one, test-pilot Tokarev V.I. reports that the functional failure in standard onboard navigators when flying at low altitudes (200–500 m) in high-latitude regions is a characteristic response of onboard navigation equipment to disturbed space weather conditions decreasing the safety of operation of military, civilian, and unmanned vehicles in the Arctic region.

Note that during extreme geomagnetic activity (GMA) due to the shift of the auroral oval to lower latitudes the risks become real also for mid-latitude technical facilities.

Thus, considering the auroras as a natural and in some cases the only available indicator of space weather conditions, it is logical to assume that the reliability of the forecast of this phenomenon correlates with the level of technosphere safety beyond the Arctic Circle. For this and other reasons, in recent decades specialists have been actively developing and improving auroral oval models built, as a rule, on the basis of long-term observations of spatial and energy characteristics of the upper ionosphere at high latitudes.

For instance, the best known model of this kind is the OVATION-Prime (OP) model [Newell et al., 2014], which is based on 21-year DMSP observations of electron and proton fluxes of different energies, takes solar wind and interplanetary magnetic field parameters, recorded at the first Lagrange point, as input, and predicts the probability of occurrence of auroras up to ~77 % [Vorobev et al., 2022b; Machol et al., 2012]. Also well-known are the predictive model NORUSCA [<http://kho.unis.no/AuroraForecast.html>; Breedveld, 2020], developed at the Norwegian Kjell Henriksen Observatory (KHO), and the diagnostic model of auroral eruptions (APM), proposed at the Polar Geophysical Institute (PGI) [Vorobjev, Yagodkina, 2005]. The latest models take a set of geomagnetic indices as input, forming an approximate geometry and location of the auroral oval as output at the time of recording of input parameters. One of the models of this type is the model developed at the University of Alaska Fairbanks, USA [<https://www.gi.alaska.edu/monitors/aurora-forecast>].

Among the means of direct ground-based observation of auroras, all-sky cameras due to their availability have become widespread [Lebedinsky, 1961; Sigernes et al., 2014]. However, the effectiveness of such observations strongly depends on environmental conditions (illumination of the sky, cloud cover, fog, etc.) and, according to the most optimistic estimates, does not exceed 35–37 %. There are also attempts to make satellite optical observations of the auroral oval, but in most cases the data is fragmentary, heterogeneous, poorly structured or unavailable.

Summing up, we note that the currently known models of auroral oval parameters are far from perfect and are in the testing phase. Thus, it is of some theoretical interest to clarify and formalize the relationship be-

tween geomagnetic field (GMF) variations and auroras. At the same time, solutions in the field of development and modernization of approaches to diagnosing properties of the upper ionosphere at high latitudes are of an applicable nature.

1. INITIAL DATA, THEIR ANALYSIS AND PREPROCESSING

In this work, the Lovozero Observatory (LOZ) is used as the main source of data on the presence of auroras, which is part of PGI and is practically the only station on the territory of the Russian Federation that continuously and for a long time has been conducting observations and recording of auroras, magnetic field variations, and other high-latitude geophysical effects, which are caused by processes in Earth's magnetosphere, ionosphere, and atmosphere. We have used nine-year data collected between 2012 and 2020 — the maximum period of openly published results of synchronous observations of auroras and GMF variations at the geographical space point considered: 67.97° N, 35.02° E (Lovozero village, Murmansk Region, Russia).

For example, the results of optical observations of auroras in the vicinity of LOZ are traditionally presented as sets of ascaplots (Figure 1) [Yagodkina, etc., 2019], published on the website of PGI since 2009 [http://pgia.ru/lang/ru/archive_pgi].

Experience indicates that the format, established since the 1970s, of presenting results of auroral observations (see Figure 1) is practically unusable in its original form in problems of mining large amounts of such data. For this reason, traditional ascaplots were previously converted into spreadsheets as in Table 1.

As a result of digitization of 1035 ascaplots for 2012–2020 (49680 episodes of 30-min observations), it has been found that only due to cloud cover the percentage of time intervals unsuitable for observing auroras significantly exceeds that of favorable periods. For instance, during the nine years of observations, complete or partial cloud cover at the zenith relative to the LOZ observatory hindered the observation of the sky in ~38.5 %

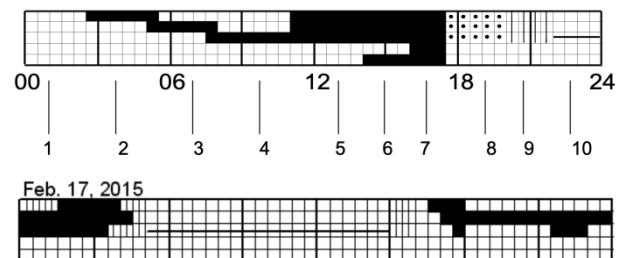


Figure 1. Format of data representation in the form of ascaplot: 1 — no aurora observed; 2 — aurora in a northern region; 3 — aurora at the zenith; 4 — aurora in the south; 5 — aurora at the zenith, in northern and southern regions; 6 — moderate aurora at the zenith; there is also a glow in the northern and southern regions; 7 — strong aurora at the zenith; there is also a glow in the northern and southern regions; 8 — partial cloud cover; 9 — complete cloud cover; 10 — no recording (a); ascaplot from the LOZ observatory on February 17, 2015 [PGI Geophysical data, 2015] (b)

of cases, while the percentage of observable fragments was ~21.85 %.

Geomagnetic variations (GMV) at the site of observation of auroras (Lovozero village) were recorded by a magnetometer of another LOZ station, located in the same geographic coordinates, but belonging to the Murmansk Department for Hydrometeorology and Environmental Monitoring (Murmansk UGMS). Geomagnetic data from the LOZ station of Murmansk UGMS is available on the SuperMag website [<https://supermag.jhuapl.edu/mag>], which in addition to collecting and storing geomagnetic data also implements some procedures for their preprocessing, for example, excludes the daily component of GMF variations, the annual trend, and the shift constant [Gjerloev, 2012]. We indicate the data thus prepared by the symbol Δ (e.g., ΔZ_{LOZ}) and examine them relative to the local magnetic coordinate system NEZ proposed by SuperMag [Gjerloev, 2012].

Table 2 in terms of the reliability theory [VorobeV et al., 2022a], presents the results of estimated completeness of time series of geomagnetic data from the LOZ station for the period under study.

Analysis of missing fragments in the geomagnetic data has shown that short-term (to 5 min) failures in the LOZ magnetometer with its subsequent recovery account for ~81.5 % of the total number of system failures and ~0.27 % of the total time of failure state. The gaps resulting from such failures were filled here by linear

interpolation without appreciable loss of data signal accuracy. The information lost due to longer episodes of failure state of the system was excluded from the general population and was not dealt with. We also excluded the values that had a clearly anomalous character against the background of their respective samples.

Thus, as a result of preprocessing for observing auroras at the zenith relative to the LOZ station, we have 9408 events that make up the set $GEN \supseteq (Cam \cap Mag)$, where Cam and Mag are subsets of all-sky camera and magnetometer observations respectively, 5430 of which correspond to the absence of auroras (NEG subset); and 3978, to their presence (POS subset), i.e. $GEN \supseteq (NEG \cup POS)$.

In addition to the LOZ magnetometer data, we propose to consider the GMA indices (*SME*, *SML*, *SMU*, *SMR*, *PCN*, etc.), published with a sampling period of at least 30 min (sampling period of ascaplots), as indicators with nonzero significance, as well as the integral aurora intensity, parameterized to the *SME* index, in the auroral zone [Newell, Gjerloev, 2011]:

$$AP \sim AP_{SME} = 0.048 \cdot SME + 0.241 \sqrt{SME}, \quad (1)$$

where *AP* is the integral aurora intensity in the auroral zone, combining auroras of four types [Newell et al., 2010].

Table 1

Fragment of a digitized ascaplot for February 17, 2015 (see Figure 1, b)

Date	UTC	Region of auroras relative to the observation				
		north	zenith	south	Aurora at the zenith, as well as in the north and south	
					zenith (moderate)	zenith (intense)
February 17, 2015	00:00	×	1	×	0	0
February 17, 2015	00:30	×	1	×	0	0
February 17, 2015	01:00	×	1	×	0	0
February 17, 2015	01:30	1	1	1	0	0
...
February 17, 2015	22:30	0	1	1	0	0
February 17, 2015	23:00	0	1	0	0	0
February 17, 2015	23:30	0	1	0	0	0

Note: 1 — aurora was observed; 0 — no aurora was observed; × — complete or partial cloud cover

Table 2

Reliability indicators of the LOZ magnetometer according to the data for 2012–2020

<i>T</i> , min	<i>T_W</i> , min	<i>T_W</i> , %	<i>T_F</i> , min	<i>T_F</i> , %	<i>N_F</i>	< <i>T_{2R}</i> >, min	< <i>T_{2F}</i> >, min
4734720	4104638	86.692	630082	13.308	632	996.97	6494.68

Note: *T* is the operating time of the LOZ magnetometer; *T_W* and *T_F* are the number of informative (total operating time) and missing (total failure time) values at the output of the LOZ magnetometer for the period *T*; <*T_{2R}*> and <*T_{2F}*> are the mean time before recovery and failure in the LOZ magnetometer respectively; *N_F* is the number of failures in the LOZ magnetometer.

2. STATISTICAL RELATIONSHIPS BETWEEN AURORAS OBSERVED IN THE VISIBLE SPECTRUM AND GEOMAGNETIC CONDITIONS

Analysis of the significance of feature variables has shown that the modulus of the first time derivative of the disturbed GMF component has the strongest connection with the objective function (observation/absence of auroras at the zenith): $|d\Delta N_{LOZ}/dt|$, $|d\Delta Z_{LOZ}/dt|$ and $|d\Delta F_{LOZ}/dt|$, where $F=(N^2+E^2+Z^2)^{1/2}$. The current situation seems to indicate that the relationship of auroras with medium-scale turbulent and wave processes in the polar ionosphere is decisive, which is reflected in these features. Especially noticeable among the significant second-order features is $|\Delta N_{LOZ}|$ obviously characterizing the relationship between auroras and the auroral electrojet intensity. Of global predictors, substorm activity indices such as SME and its derivative AP_{SME} demonstrate a quite close correlation with observations of auroras.

Figure 2 illustrates distributions of the most significant local features. We can see that in terms of the general population (GEN) the statistics of 30-min averaged values corresponds to lognormal law (2) and generalized Pareto distribution (3) characterizing heavy tails defined within ~ 85.8 percentile (*a*).

$$PDF(x, s) = \frac{1}{sx\sqrt{2\pi}} \exp\left(-\frac{\log^2 x}{2s^2}\right), \quad (2)$$

$$PDF(x, c) = (1+cx)^{-\frac{1}{c}}, \quad (3)$$

where s and c are shape parameters.

During auroral observations (POS subset data), the features exhibit a similar statistical behavior. The absolute heavy-tail boundaries (6.3 ± 1.1 nT/min) remain unchanged such that the boundaries in relative values decrease from ~ 85.8 to ~ 67.5 percentile (Figure 2, *a, b*). Without auroras (NEG data), distribution of the features becomes almost homogeneous and is described exclusively by a lognormal law (Figure 2, *c*).

Moreover, the statistics of features in the absence of auroras have maximum asymmetry and kurtosis, which correspond to the heaviest tails, whereas the asymmetry and kurtosis of the same feature in the POS subset are minimum. The current situation indicates that in the case of NEG samples the features are most densely concentrated in the vicinity of the mean value; therefore, even with small increments of the feature variables the probability of observing auroras increases sharply.

From this we can deduce that the geomagnetic conditions when there are no auroras observed are more deterministic than when auroras occur, i.e. the conditions are less uncertain. In other words, extreme GMDs can in fact ensure observations of auroras, yet the fact of observing auroras does not guarantee strong GMF variations and is statistically due to the disturbances only in $\sim 1/3$ of cases (Figure 2, *b*).

Figure 3 displays auroras observed at the zenith of the LOZ station at different GMA levels relative to the boundary of heavy-tail detection (see Figure 2, *b*). Thus, at $|d\Delta N_{LOZ}/dt|, |d\Delta Z_{LOZ}/dt|, |d\Delta F_{LOZ}/dt| \ll 5$ nT/min, faint northward auroras are observed (Figure 3, *a*); at $|d\Delta N_{LOZ}/dt|, |d\Delta Z_{LOZ}/dt|, |d\Delta F_{LOZ}/dt| \sim 7$ nT/min, a characteristic arc is likely to exist (Figure 3, *b*), which is transformed into an active arc, spiral or vortex at $|d\Delta N_{LOZ}/dt|, |d\Delta Z_{LOZ}/dt|, |d\Delta F_{LOZ}/dt| \gg 7$ nT/min (Figure 3, *c*).

Interestingly, the statistics of the most significant global features (Figure 4) structurally repeats the distribution of local predictors, and the distribution of AP_{SME} does not contradict the known interpretation of AP in the auroral zone: at $AP < 20$ GW, a weak or faint aurora is observed; at $20 \leq AP \leq 50$ GW, an aurora can be seen, but at a short distance from it; at $50 < AP \leq 100$ GW, an aurora can be seen with the naked eye; $AP > 100$ GW corresponds to extreme auroral activity and a significant expansion of the auroral oval [Vorobev et al., 2022b]

3. SYNTHESIS AND VERIFICATION OF DIAGNOSIS MODELS

As follows from Figures 2, 4, the probability of observing auroras is maximum when features exceed some reference values: $AP_{SME} > 70$ GW; $|d\Delta N_{LOZ}/dt|, |d\Delta Z_{LOZ}/dt|, |d\Delta F_{LOZ}/dt| > 13$ nT/min. In general, the fulfillment of this condition at the interval of probable observation of auroras (Figure 5) makes it possible to effectively identify their presence, for instance, under conditions of complete or partial cloud cover.

However, the same Figures suggest that most auroras recorded correspond to the predictor values lower than the reference ones, thereby adding uncertainty to the diagnostic result and affecting the quality of the results thus obtained. A way out of this situation can be the use of more advanced binary classification methods such as the Bayesian classifier, the logistic regression, the random forest method, etc.

3.1. Diagnostics of aurora presence based on Bayesian inference

Let us analyze the basic approach to identifying auroras from ground magnetic station data, using Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (4)$$

where $P(A)$ is an a priori probability of hypothesis A or an a priori distribution; $P(A|B)$ is a probability of hypothesis A upon occurrence of event B (a posteriori probability); $P(B|A)$ is a probability of occurrence of event B if hypothesis A is true; $P(B)$ is the total probability of occurrence of event B , defined according to Expression (5).

$$P(B) = \sum_{i=1}^N P(B|A_i)P(A_i), \quad (5)$$

where the probabilities under the summation sign are known or allow an experimental estimate.

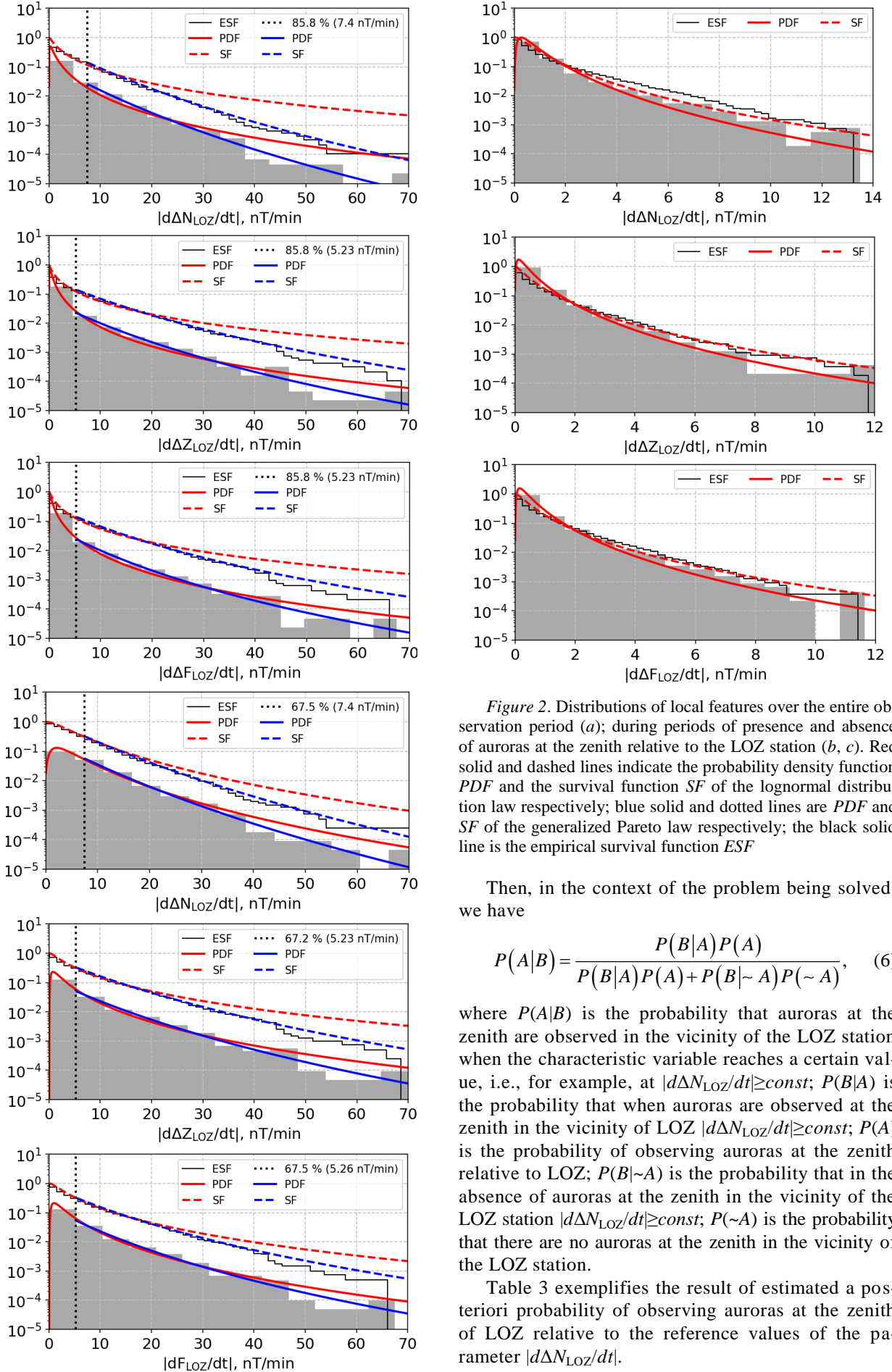


Figure 2. Distributions of local features over the entire observation period (a); during periods of presence and absence of auroras at the zenith relative to the LOZ station (b, c). Red solid and dashed lines indicate the probability density function PDF and the survival function SF of the lognormal distribution law respectively; blue solid and dotted lines are PDF and SF of the generalized Pareto law respectively; the black solid line is the empirical survival function ESF

Then, in the context of the problem being solved, we have

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}, \quad (6)$$

where $P(A|B)$ is the probability that auroras at the zenith are observed in the vicinity of the LOZ station when the characteristic variable reaches a certain value, i.e., for example, at $|d\Delta N_{LOZ}/dt| \geq const$; $P(B|A)$ is the probability that when auroras are observed at the zenith in the vicinity of LOZ $|d\Delta N_{LOZ}/dt| \geq const$; $P(A)$ is the probability of observing auroras at the zenith relative to LOZ; $P(B|\sim A)$ is the probability that in the absence of auroras at the zenith in the vicinity of the LOZ station $|d\Delta N_{LOZ}/dt| \geq const$; $P(\sim A)$ is the probability that there are no auroras at the zenith in the vicinity of the LOZ station.

Table 3 exemplifies the result of estimated a posteriori probability of observing auroras at the zenith of LOZ relative to the reference values of the parameter $|d\Delta N_{LOZ}/dt|$.

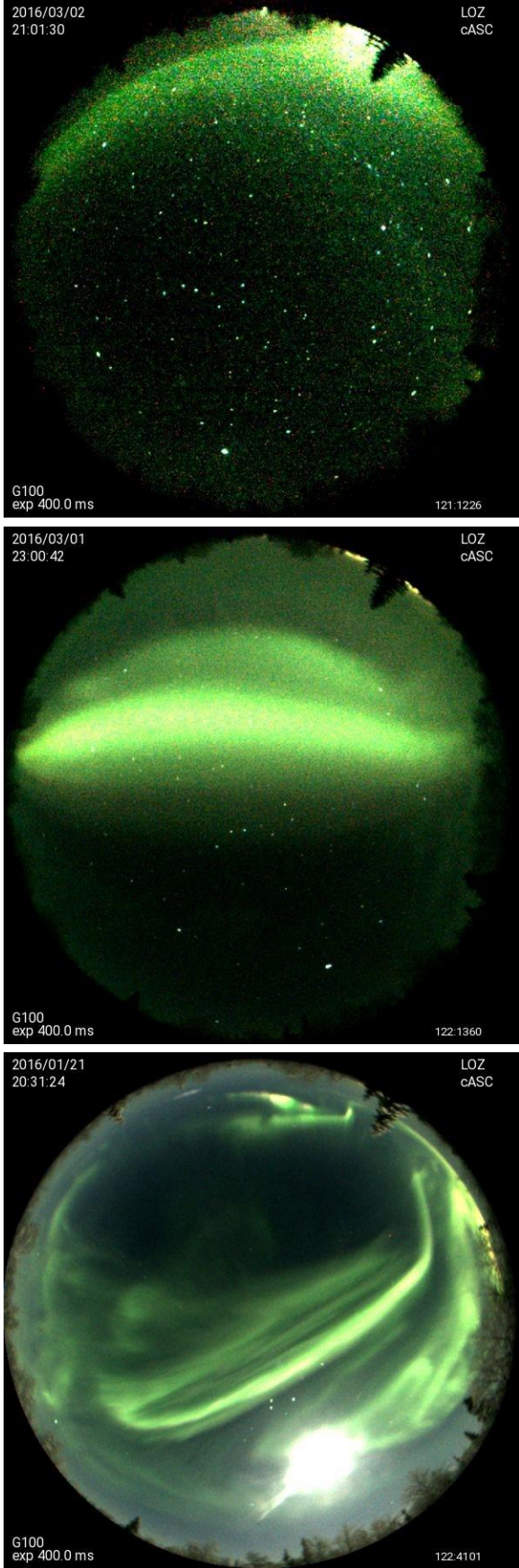


Figure 3. All-sky camera images from the LOZ observatory at $(|d\Delta N_{LOZ}/dt|, |d\Delta Z_{LOZ}/dt|, |d\Delta F_{LOZ}/dt|) \ll 5$ nT/min (a); $(|d\Delta N_{LOZ}/dt|, |d\Delta Z_{LOZ}/dt|, |d\Delta F_{LOZ}/dt|) \sim 7$ nT/min (b); $(|d\Delta N_{LOZ}/dt|, |d\Delta Z_{LOZ}/dt|, |d\Delta F_{LOZ}/dt|) \gg 7$ nT/min (c)

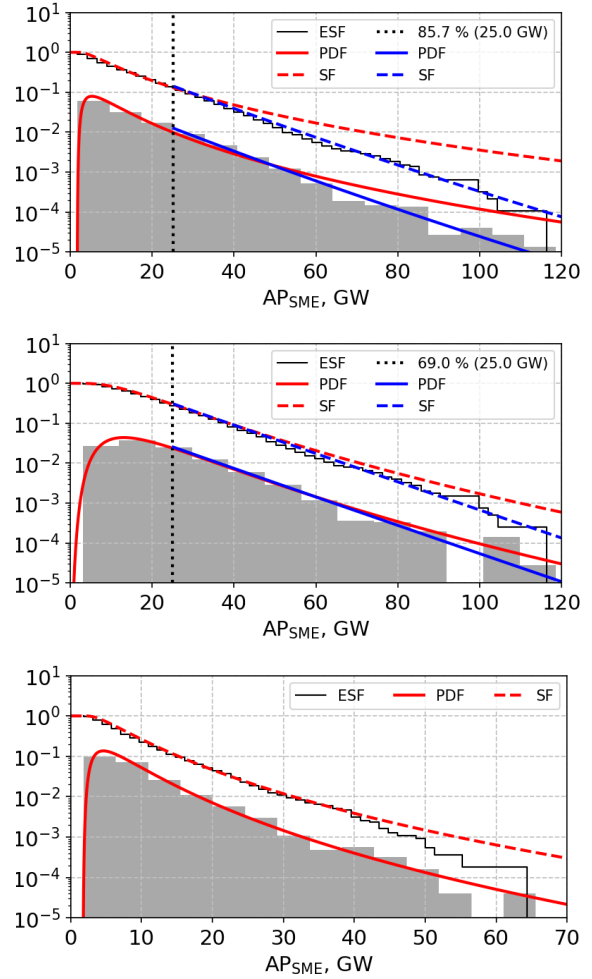


Figure 4. Statistics parameterized to the SME index of the integral aurora intensity in the auroral zone: a — for the entire observation period; b, c — in the presence and absence of auroras at the zenith relative to the LOZ station respectively

Figure 6 exhibits the functional dependence of the a posteriori probability of observing auroras on the most significant predictors. Such dependence has a clear logarithmic character and can generally be approximated by function (7). In this case, the significance level of the feature is inversely proportional to the discrepancy between the empirical data and the approximating function.

$$P(A|B) \approx P(X) = a \ln(bX + c), \quad (7)$$

where X is a feature variable; a, b, c are its related shape parameters; for $|dN_{LOZ}/dt|$ $a=7.04$; $b=1.32 \cdot 10^5$; $c=-1.14 \cdot 10^5$.

3.2. Diagnostics of aurora presence by machine learning methods

A preliminary assessment of the quality metrics of several classical approaches to binary classification in the context of the problem being solved reveals an advantage of the random forest method. However, along with the best diagnostic accuracy, this method is practically uninterpretable, i.e., if implemented, the diagnosis model for the end user will be a "black box" generating the state of the target function, without indicating in any way the mechanisms that determine the diagnostic result.

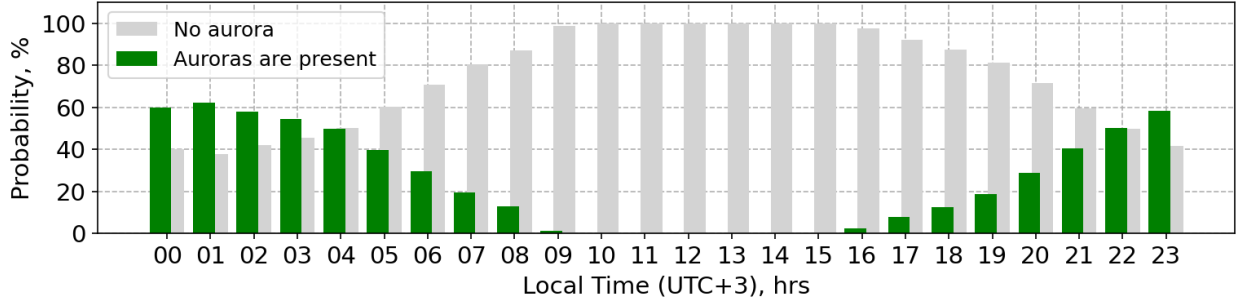


Figure 5. Daily variation in the probability of observing auroras at the zenith of the LOZ station, calculated from ascaplots, using data published by PGI for 2012–2020.

Table 3

Estimated a posteriori probability of the presence of auroras at the zenith relative to the LOZ station

$ d\Delta N_{\text{Loz}}/dt $, nT/min	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 6	≥ 7	≥ 8	≥ 9	≥ 10
$P(B A)$, %	91.23	80.22	68.9	59.55	50.73	42.28	35.14	29.49	24.66	20.66
$P(B \sim A)$, %	29.78	12.06	6.67	3.96	2.56	1.62	1.07	0.64	0.35	0.22
$P(A B)$, %	69.18	82.97	88.33	91.68	93.56	95.03	96.02	97.1	98.1	98.56

Note: $P(A)=3978/9408\approx 42.28\%$; $P(\sim A)=5430/9408\approx 57.72\%$.

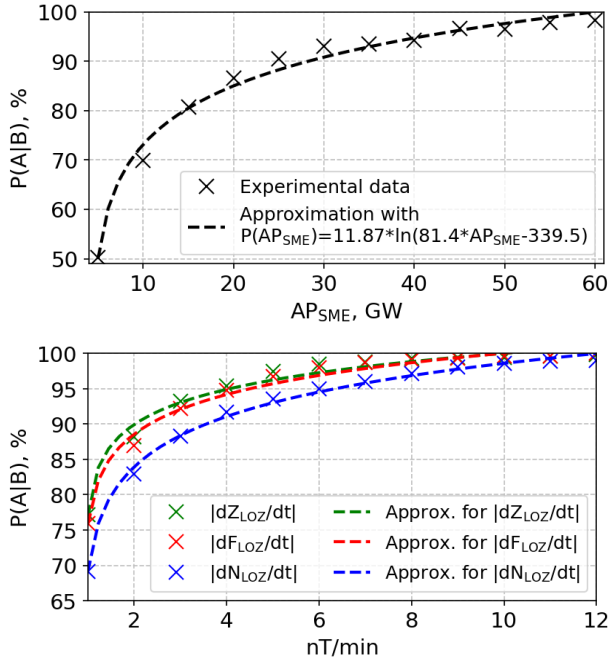


Figure 6. Type of the dependence of a posteriori probability of observing auroras at the zenith relative to the LOZ station on the most significant global (a) and local (b) feature variables

Assessing the significance of predictors according to the Gini criterion [Witlox, 2017] for the random forest model made it possible to rank feature variables independently of the previously obtained results and to identify parameters contributing to the reliability of diagnostic results. The result obtained correlates well with the already available data and mainly for the features that have the highest significance.

Thus, when selecting the nine most significant predictors that are not related to each other by linear dependence (Table 4) and optimizing hyperparameters of

the model (the number of trees is 400, the number of random features for choosing trees splitting is $\log_2(M)+1\approx 4$, where $M=9$ is the number of features of the model [Mantas et al., 2019], the minimum number of objects in leaves is 3, the maximum depth of the tree is 8), the accuracy of the diagnostics of aurora presence was at least 86.3 % (Tables 4, 5).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \quad (8)$$

where TP and TN are true-positive and true-negative diagnoses respectively; FP and FN are false-positive and false-negative ones.

The presence of auroras on the basis of four most significant features (see Table 4) can be identified up to 85.7 %; the training time of the model in this case decreases ~ 4.1 times and is 12.6 s. If exclusively SME and SMR are utilized as input parameters, the percentage of model errors does not exceed 19.7 % (i.e. $\text{Accuracy} \geq 80.3\%$), and the result may be relevant for the entire set of points located at the geomagnetic latitude of the LOZ station at night.

4. DISCUSSION

The studies conducted strongly suggest that it is possible to effectively identify the presence of auroras by machine learning methods, statistical and intelligent analyses of geomagnetic data from ground sources. At the same time, the results are of practical significance for problems of organizing decision-making support for the identification of auroras by non-automated analysis of keograms and/or all-sky camera data. In the future, the scope of application of the results can be expanded to near real-time diagnostics of the upper ionosphere conditions at high latitudes, assessments of the risks of failure in HF radio communication systems and extreme errors in global navigation satellite systems operated in the Arctic.

Table 4

Quality metrics of diagnosis models for different sets of input features		Metrics	
Input features		<i>AUC</i> *	<i>Accuracy</i> **
$ d\Delta N_{LOZ}/dt $, $ d\Delta E_{LOZ}/dt $, $ d\Delta Z_{LOZ}/dt $, $ d\Delta F_{LOZ}/dt $, $ \Delta N_{LOZ} $, $ \Delta E_{LOZ} $, $ \Delta Z_{LOZ} $, <i>SME</i> , <i>SMR</i>		0.927	0.863
$ d\Delta N_{LOZ}/dt $, $ d\Delta Z_{LOZ}/dt $, $ d\Delta F_{LOZ}/dt $, $ \Delta N_{LOZ} $		0.924	0.857
<i>SME</i> , <i>SMR</i>		0.879	0.803

Note: **AUC* (Area Under Curve) — the area under the ROC curve [Hand, Till, 2001] and the axis of the percentage of false positive classifications; ** *Accuracy* is determined by Expression (8).

Table 5

Estimated quality of diagnostics of aurora presence in test data (25 % of GEN), when using nine most significant feature variables as input

Aurora at the zenith	Observed	Not observed
Aurora diagnosis		
positive	1227	140
negative	182	803

Due to the initially rather weak features, the accuracy of the obtained models is relatively low, ~86 %. In this regard, the issues that especially require further research are those relating to the identification of more complex and strong synthetic predictors, whose existence is clearly evident from the results of the analysis of the main components [Jolliffe, 2002]. It also makes sense to address the questions of systematization and complex processing of data from several high-latitude magnetometers and all-sky cameras in a small subregion (for example, a subregion bounded by the observatories Lovozero, Apatity, and Verkhnetulomsky).

CONCLUSIONS

The existing variety of approaches to monitoring space weather and geophysical parameters in the auroral oval do not solve the problem of effective forecasting and diagnostics of auroras as a special state of the upper ionosphere at high latitudes. The format of presenting data from all-sky cameras in the form of ascaplots, unchanged since the 1970s, has practically lost its relevance today and needs in-depth modernization, for example, by organizing automated data separation, reducing the sampling increment, and modifying the approach to classifying observable events.

The recursive feature elimination method [Kuhn, Johnson, 2019] and the estimated mutual information criterion [Baudot et al., 2019] indicate that the first time derivative of the northward and vertical components of GMF variations most closely related to the occurrence of auroras in the optical range, which is likely suggestive of an essential relationship of auroras with medium-scale turbulent and wave processes in the polar ionosphere. Of the significant second-order features, $|\Delta N_{LOZ}|$ and the *SME* index stand out as characterizing the relationship of auroras with the auroral electrojet intensity and the substorm activity level in general.

Statistical analysis of geomagnetic data for nine years (2012–2020) suggests that the geomagnetic condi-

tions in the absence of auroras have a lower degree of uncertainty than when auroras occur. In other words, extreme GMDs cause auroras in the optical range, yet the fact of observing auroras does not guarantee the occurrence of strong GMF variations and is statistically due to the disturbances only in ~1/3 of cases.

The dependence of the a posteriori probability of observing auroras on observable features has a clear logarithmic character and can generally be approximated by a function of the type $P(A|B) \approx P(X) = a \ln(bX+c)$, where X is a feature variable; a , b , c are its related shape parameters (for $|dN_{LOZ}/dt|$ $a=7.04$; $b=1.32 \times 10^5$; $c=-1.14 \times 10^5$). In this case, the degree of significance of the feature is inversely proportional to the discrepancy between empirical data and the approximating function.

The accuracy of local identification of auroras from geomagnetic data based on the random forest method and a number of the most significant feature variables is ~86 %. The diagnosis model based on global geomagnetic indices has the expected lower accuracy of ~80.3 %; however, the data obtained from it can be used to verify known global diagnosis models of the auroral oval with a similar set of input parameters (for example, the auroral precipitation model [Vorobjev, Yagodkina, 2005]).

The study was financially supported by RSF (Project No. 21-77-30010).

We are grateful to the reviewers for careful analysis of the work and a large number of constructive comments, as well as to V.I. Tokarev, a test-pilot at the Yu.A. Gagarin Cosmonaut Training Center, for information about the unique experience of piloting aircraft in the Arctic airspace during periods of extreme geomagnetic activity.

REFERENCES

Baudot P., Tapia M., Bennequin D., Goillard J.-M. Topological Information Data Analysis. *Entropy*. 2019, vol. 21, iss.9, p. 869. DOI: [10.3390/e21090869](https://doi.org/10.3390/e21090869).

- Breedveld M.J. *Predicting the Auroral Oval Boundaries by Means of Polar Operational Environmental Satellite Particle Precipitation Data*. Master Thesis. Arctic University of Norway. June 2020.
- Gjerloev J.W. The SuperMAG data processing technique. *J. Geophys. Res.* 2012, vol. 117, iss. A9, p. A09213. DOI: [10.1029/2012JA017683](https://doi.org/10.1029/2012JA017683).
- Hand D.J., Till R.J. A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning*. 2001, vol. 45, pp. 171–186. DOI: [10.1023/A:1010920819831](https://doi.org/10.1023/A:1010920819831).
- Jolliffe I.T. *Principal Component Analysis*. Ser.: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p.
- Kuhn M., Johnson K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019, 298 p.
- Lebedinsky A.I. Synchronous auroral registration by all-sky camera C-180 and patrol spectrograph C-180-S. *Ann. Intern. Geophys. Year*. 1961, vol. XI.
- Machol J.L., Green J.C., Redmon R.J., Viereck R.A., Newell P.T. Evaluation of OVATION as a forecast model for visible aurorae. *Space Weather*. 2012, vol. 10, iss. 3, p. S03005. DOI: [10.1029/2011SW000746](https://doi.org/10.1029/2011SW000746).
- Mantas C.J., Castellano J.G., Moral-García S., Abellán J. A comparison of random forest based algorithms: random credal random forest versus oblique random forest. *Soft Computing*. 2019, vol. 23, pp. 10739–10754. DOI: [10.1007/s00500-018-3628-5](https://doi.org/10.1007/s00500-018-3628-5).
- Newell P.T., Gjerloev J.W. Substorm and magnetosphere characteristic scales inferred from the SuperMAG auroral electrojet indices. *J. Geophys. Res.* 2011, vol. 116, iss. A12, p. A12232. DOI: [10.1029/2011JA016936](https://doi.org/10.1029/2011JA016936).
- Newell P.T., Sotirelis T., Wing S. Seasonal variations in diffuse, monoenergetic, and broadband aurora. *J. Geophys. Res.* 2010, vol. 115, iss. A3, p. A03216. DOI: [10.1029/2009JA014805](https://doi.org/10.1029/2009JA014805).
- Newell P.T., Liou K., Zhang Y., Sotirelis T., Paxton L.J., Mitchell E.J. OVATION Prime-2013: Extension of auroral precipitation model to higher disturbance levels. *Space Weather*. 2014, vol. 12, iss. 6, pp. 368–379. DOI: [10.1002/2014SW001056](https://doi.org/10.1002/2014SW001056).
- PGI Geophysical data. January, February, March 2015 / Ed. V. Vorobjev. Murmansk, Apatity: PGI KSC RAS. 2015.
- Pilipenko V.A. Space weather impact on ground-based technological systems. *Solar-Terr. Phys.* 2021, vol. 7, iss. 3, pp. 68–104. DOI: [10.12737/stp-73202106](https://doi.org/10.12737/stp-73202106).
- Ptitsyna N.G., Tyasto M.I., Kasinsky V.V., Lyakhov N.N. Influence of space weather on technical systems: failures of railway equipment during geomagnetic storms. *Solar-Terr. Phys.* 2008, No. 12-2 (125), pp. 360. (In Russian).
- Sigernes F., Holmen S. E., Biles D., Bjørklund H., Chen X., Dyrland M., Lorentzen D.A., Baddeley L., et al. Auroral all-sky camera calibration. *Geoscientific Instrumentation, Methods and Data Systems*. 2014, vol. 3, iss. 2, pp. 241–245. DOI: [10.5194/gi-3-241-2014](https://doi.org/10.5194/gi-3-241-2014).
- Sokolova O.N., Sakharov Ya.A., Gritsutenko S.S., Korovkin N.V. Algorithm for analyzing the stability of power systems to geomagnetic storms. *News of the Russian Academy of Sciences. Energy*. 2019, no. 5, pp. 33–52. DOI: [10.1134/S0002331019050145](https://doi.org/10.1134/S0002331019050145). (In Russian).
- Soloviev A.A., Sidorov R.V., Oshchenko A.A., Zaitsev A.N. On the need for accurate monitoring of the geomagnetic field during directional drilling in the Russian Arctic. *Izvestiya. Physics of the Solid Earth*. 2022, vol. 58, pp. 420–434. DOI: [10.1134/S1069351322020124](https://doi.org/10.1134/S1069351322020124).
- Vorobev A., Soloviev A., Pilipenko V., Vorobeva G., Sakharov Y. An approach to diagnostics of geomagnetically induced currents based on ground magnetometers data. *App. Sci.* 2022a, vol. 12, iss. 3, pp. 1522. DOI: [10.3390/app12031522](https://doi.org/10.3390/app12031522).
- Vorobev A.V., Soloviev A.A., Pilipenko V.A., Vorobeva G.R. Interactive Computer model for aurora forecast and analysis. *Solar-Terr. Phys.* 2022b, vol. 8, no 2, pp. 84–90. DOI: [10.12737/stp-82202213](https://doi.org/10.12737/stp-82202213).
- Vorobjev V.G., Yagodkina O.I. Effect of magnetic activity on the global distribution of auroral precipitation zones. *Geomagnetism and Aeronomy*. 2005, vol. 45, pp. 438–444.
- Witlox F. Gini Coefficient. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*. 2017. DOI: [10.1002/9781118786352.wbieg0855](https://doi.org/10.1002/9781118786352.wbieg0855).
- Yagodkina O.I., Vorobyov V.G., Shekunova E.S. Observations of auroras over the Kola Peninsula. *Proc. Kola Scientific Center of the Russian Academy of Sciences*. 2019, vol. 10, no. 8-5, pp. 43–55. DOI: [10.25702/KSC.2307-5252.2019.10.8](https://doi.org/10.25702/KSC.2307-5252.2019.10.8). (In Russian).
- Yasyukevich Y., Astafyeva E., Padokhin A. Ivanova V., Syrovatskii S., Podlesnyi A. The 6 September 2017 X-class solar flares and their impacts on the ionosphere, GNSS, and HF radio wave propagation. *Space Weather*. 2018, vol. 16, iss. 8, pp. 1013–1027. DOI: [10.1029/2018SW001932](https://doi.org/10.1029/2018SW001932).
- Yasyukevich Y., Vasilyev R., Ratovsky K. Small-scale ionospheric irregularities of auroral origin at mid-latitudes during the 22 June 2015 magnetic storm and their effect on GPS positioning. *Remote Sensing*. 2020, vol. 12, no 10, p. 1579. DOI: [10.3390/rs12101579](https://doi.org/10.3390/rs12101579).
- Zakharov V.I., Chernyshov A.A., Miloh V., Jin Ya. Influence of the ionosphere on the parameters of GPS navigation signals during a geomagnetic substorm. *Space Res.* 2020, vol. 60, no. 6, pp. 769–782. DOI: [10.7868/S0023420616010143](https://doi.org/10.7868/S0023420616010143). (In Russian).
- Original Russian version: Vorobev A.V., Soloviev A.A., Pilipenko V.A., Vorobeva G.R., Gainetdinova A.A., Lapin A.N., Belakhovsky V.B., Roldugin A.V., published in *Solnechno-zemnaya fizika*. 2023. Vol. 9. Iss. 2. P. 26–34. DOI: [10.12737/szf-92202303](https://doi.org/10.12737/szf-92202303). © 2023 INFRA-M Academic Publishing House (Nauchno-Izdatelskii Tsentr INFRA-M)
- How to cite this article*
- Vorobev A.V., Soloviev A.A., Pilipenko V.A., Vorobeva G.R., Gainetdinova A.A., Lapin A.N., Belakhovsky V.B., Roldugin A.V. Local diagnostics of aurora presence based on intelligent analysis of geomagnetic data. *Solar-Terrestrial Physics*. 2023. Vol. 9. Iss. 2. P. 22–30. DOI: [10.12737/stp-92202303](https://doi.org/10.12737/stp-92202303).