applied sciences

MDPI

Article

Artificial Neural Networks for Determining the Empirical Relationship between Meteorological Parameters and High-Level Cloud Characteristics

Olesia Kuchinskaia [1,*], Maxim Penzin [1], Iurii Bordulev [1], Vadim Kostyukhin [1], Ilia Bryukhanov [1,2,3], Evgeny Ni [1,2], Anton Doroshkevich [1,2], Ivan Zhivotenyuk [1,2], Sergei Volkov [3] and Ignatii Samokhvalov [1,2]

[1] Laboratory for Analysis of High Energy Physics Data, Faculty of Physics, National Research Tomsk State University, 634050 Tomsk, Russia; penzin.maksim@gmail.com (M.P.); bordulev@gmail.com (I.B.); vadim.kostyukhin@cern.ch (V.K.); plyton@mail.tsu.ru (I.B.); niev@mail.tsu.ru (E.N.); adoro@mail.tsu.ru (A.D.); guitarplayer@mail.tsu.ru (I.Z.); lidar@mail.tsu.ru (I.S.)
[2] Department of Optoelectronic Systems and Remote Sensing, Faculty of Radiophysics, National Research Tomsk State University, 634050 Tomsk, Russia
[3] Center of Laser Atmosphere Sensing, V.E. Zuev Institute of Atmospheric Optics of the Siberian Branch of the Russian Academy of Sciences, 634055 Tomsk, Russia; snvolk@iao.ru
* Correspondence: olesia.kuchinskaia@cern.ch; Tel.: +7-923-421-77-13

Abstract: The special features of the applicability of artificial neural networks to the task of identifying relationships between meteorological parameters of the atmosphere and optical and geometric characteristics of high-level clouds (HLCs) containing ice crystals are investigated. The existing models describing such relationships do not take into account a number of atmospheric effects, in particular, the orientation of crystalline ice particles due to the simplified physical description of the medium, or within the framework of these models, accounting for such dependencies becomes a highly nontrivial task. Neural networks are able to take into account the complex interaction of meteorological parameters with each other, as well as reconstruct almost any dependence of the HLC characteristics on these parameters. In the process of prototyping the software product, the greatest difficulty was in determining the network architecture, the loss function, and the method of supplying the input parameters (attributes). Each of these problems affected the most important issue of neural networks—the overtraining problem, which occurs when the neural network stops summarizing data and starts to tune to them. Dependence on meteorological parameters was revealed for the following quantities: the altitude of the cloud center; elements m_{22} and m_{44} of the backscattering phase matrix (BSPM); and the m_{33} element of BSPM requires further investigation and expansion of the analyzed dataset. Significantly, the result is not affected by the compression method chosen to reduce the data dimensionality. In almost all cases, the random forest method gave a better result than a simple multilayer perceptron.

Keywords: big data; artificial neural networks; simple multilayer perceptron; random forest method; cross-validation; atmosphere; high-level clouds; horizontally oriented ice particles; polarization lidar; ERA5 reanalysis

Citation: Kuchinskaia, O.; Penzin, M.; Bordulev, I.; Kostyukhin, V.; Bryukhanov, I.; Ni, E.; Doroshkevich, A.; Zhivotenyuk, I.; Volkov, S.; Samokhvalov, I. Artificial Neural Networks for Determining the Empirical Relationship between Meteorological Parameters and High-Level Cloud Characteristics. *Appl. Sci.* **2024**, *14*, 1782. https://doi.org/10.3390/app14051782

Academic Editors: Joao Carlos Andrade dos Santos and André Ribeiro Da Fonseca

Received: 8 January 2024
Revised: 6 February 2024
Accepted: 19 February 2024
Published: 22 February 2024

1. Introduction

Climate change has become one of the most significant global challenges of our time. The application of machine learning (ML) methods provides the ability to not only process and analyze huge amounts of information but also to reveal patterns that are inaccessible using other tools. This enables scientists to better understand the impact of a variety of factors on climate processes and make more accurate forecasts. Modern hardware and software tools allow us to deepen the understanding of atmospheric processes [1,2] and improve forecasting in the field of environmental engineering [3–5], and they are even

used to optimize power systems [6]. Cloudiness is the most important regulator of the Earth's radiation budget. However, a wide range of questions related to the evaluation of its influence on the amount of solar energy reaching the Earth's surface remains unresolved. High-level clouds (HLCs) are known for their large horizontal extent, which can be up to thousands of kilometers [7,8]. For this reason, they significantly affect the regional and global radiation budget and climate due to the effects of attenuation and reflection of radiation [9,10].

Neural networks (NNs) can be a powerful tool in solving climate problems, but they are not a universal solution for all tasks and must be used in combination with other research and data analysis techniques to achieve the best results. In order for a NN to be able to solve a problem, it needs to be properly trained on a sufficiently large amount of high-quality data, and it needs to take into account all factors affecting the problem being solved. When using NNs in climate research, it should be recognized that they can be ineffective when important climate factors have not been included in the initial dataset, which can lead to incorrect conclusions.

Thus, in order to achieve the best results, it is necessary to use multimodal studies that include different methods of data investigation and analysis. The present research describes this approach: we investigated the special features of the NN applicability to the problem of reconstructing the relationship between meteorological parameters and optical characteristics of HLCs. The existing atmospheric models estimate the microstructure of HLCs quite roughly: a real ice particle is replaced by an equivalent sphere with a certain value of the effective radius, which negatively affects the accuracy of weather and climate forecasts [11,12]. The special feature of our approach is the use of the set of atmosphere lidar sensing data from 2009 to the present, which combines more than 3 thousand series of measurements of atmospheric parameters and is systematically expanded. The lidar was developed and built at the National Research Tomsk State University (NR TSU) [13]. Lidar methods are the most promising in solving the tasks of operative control and monitoring of the atmosphere state since they allow vertical profiles of optical, microphysical, and meteorological parameters to be determined in real time [14–17].

The materials and methods are discussed in Section 2. Section 2.1 contains basic information about the high-altitude matrix polarization lidar (HAMPL). Section 2.2 describes sources of information on meteorological conditions at the altitudes of clouds registered by the lidar. Section 2.3 provides basic information about the software product developed for processing the lidar meteorological dataset based on machine learning methods. Section 3 collects a description of the results of the work. Section 3.1 gives a preliminary analysis of the data. Section 3.2 describes the implementation of data dimensionality reduction. Section 3.3 is devoted to the implementation of the estimation of the HLC detection altitude. Section 3.4 presents the results of the HLC upper- and lower-boundary altitudes, and Section 3.5 presents the estimation of elements of the HLC backscattering phase matrix (BSPM). Section 4 discusses the results of the paper. The conclusion is given in Section 5.

2. Materials and Methods

2.1. High-Altitude Matrix Polarization Lidar (HAMPL)

A block diagram of the HAMPL and the issues of detecting the preferred horizontal orientation of ice particles in HLCs are described in [18]. The lidar is oriented vertically in the zenith direction (Figure 1). Measurements are performed at any time of the day provided that there is no precipitation, strong wind, or low clouds. A Nd:YAG laser Lotis TII LS-2137U with an operating wavelength of 532 nm, a pulse energy of up to 400 mJ, and a pulse repetition rate of 10 Hz is used as an optical radiation source. A Cassegrain mirror lens with a primary mirror diameter of 0.5 m and a focal length of 5 m is used as a receiving antenna. A Wollaston prism, which divides received backscattered radiation into two mutually orthogonally polarized beams, is installed on the output of the receiving optical channel of the lidar. These radiation beams are recorded by two Hamamatsu H5783P

photomultiplier tubes (PMTs) operating in the photon-counting mode with time strobing of the signal, which provides the lidar altitude resolution from 37.5 to 150 m.

Figure 1. Photographs of the NR TSU HAMPL: general view of the transceiver part (4th floor of the building), view on the roof of the building, and view from the roof of the building (in the hatch down to the 4th and 5th floors of the building).

To suppress active backscatter noise from the near-field zone of the lidar (up to 3 km), electro-optical shutters (EOSs) are installed in front of the PMTs. The EOSs are based on a potassium dideuterium phosphate (DKDP) crystal with a maximum trigger frequency of 100 Hz, a high-voltage pulse duration of 1–1000 μs, and a high-voltage pulse delay relative to the trigger pulse of 1–100 μs. The use of EOS allows the characteristics to remain linear even during lidar operation in the daytime at the maximum energy of the sensing pulse. Radiation pulses with four different polarization states (three linear and one circular) are sent alternately into the atmosphere, for each of which the polarization state of backscattered radiation described by the Stokes vector parameter is determined. Thus, 16 intensity profiles are measured in each sensing cycle, from which 16 BSPM elements are calculated. BSPM is a 4 × 4 matrix with m_{ij} (i, j = 1, 2, 3, 4) elements, which is the operator for converting the Stokes vector parameter describing the polarization state of probing radiation into the Stokes vector parameter of radiation backscattered by the cloud. In addition, other important characteristics of clouds are determined based on the analysis of lidar measurement data: optical—scattering ratio and optical thickness—and geometric—altitudes of lower and upper boundaries and vertical power.

The parallel accumulation of signals allows a long continuous session of HLC sensing to be performed and the most interesting time intervals to be selected for processing. The application of such a procedure allowed us to establish that the same extended cloud can contain both specular (containing horizontally oriented ice particles) and nonspecular local areas and to estimate their sizes.

2.2. Meteorological Conditions at the Altitudes of Clouds Registered by the Lidar

The study of cloud characteristics with the lidar requires knowledge of the meteorological parameter values at their altitudes. The most reliable source of vertical profiles of meteorological values is the radiosonde observation. Such measurements are not performed systematically directly in Tomsk. The nearest aerological stations to the HAMPL location are located in Kolpashevo (WMO 29231) and Novosibirsk (WMO 29634) [19]. Data from these stations are available online [20]. In spite of the distance of about 230–250 km of each of them from Tomsk, the meteorological conditions at the altitudes of HLC formation according to the data of these stations are usually close. Nevertheless, differences are sometimes observed. In addition, according to the standard procedure of the World Meteo-

rological Organization, these stations launch radiosondes only twice a day. Thus, when processing the results of lidar measurements, it is necessary to choose the data considering the above-mentioned aerological stations and the time of sounding (morning or evening).

In connection with these circumstances, we have previously evaluated [18,21] the possibility of using the ERA5 reanalysis data [22] from the European Centre for Medium-Range Weather Forecasts (ECMWF) to interpret the results of lidar measurements. Their verification was performed on the basis of comparison with the data of radiosonde observation at five aerological stations within 500 km from Tomsk for each day for five consecutive years. The correctness of using the vertical profiles of some meteorological variables in the range of altitudes of HLC formation was shown. Therefore, we have recently used the ERA5 reanalysis as an alternative source of meteorological information.

2.3. Machine Learning Techniques

In the framework of the present work, the applicability of artificial neural networks to the task of reconstructing the relationships between meteorological parameters of the atmosphere and optical and geometric characteristics of HLCs was investigated. Neural networks are able to take into account the complex interaction of meteorological parameters with each other, as well as to reconstruct practically any dependence of the HLC characteristics on these parameters. In the process of prototyping the software product (SP), the greatest difficulty was determining the network architecture, the loss function, and the method of supplying the input parameters (features). Each of these problems affected the most important problem of neural networks—the overtraining problem, which occurs when the neural network stops summarizing data and starts to tune to them.

When choosing a neural network architecture and one of the possible machine learning models, we were guided by a practical heuristic: the number of parameters used to tune the model should be smaller than the set of data used for training. When this condition is met, the neural network begins to generalize the data rather than memorize it. This also allows us to limit the set of potential architectures that can be used for training. For this purpose, a preliminary analysis was made of lidar measurements of HLC BSPM for the period 2016–2023 (2009–2015 differed in the way HLCs were measured, so the data require long and more complex processing and will be available soon). The dataset was summarized into a database (DB), which contains information on 1177 measurement series (the duration of one measurement series was about 17 min). In 312 cases of these, HLCs were detected. In addition, in another 172 series, HLCs were also registered, but the lidar signal from them was insufficient for statistical processing. These data were used to verify the prediction by the neural network of the altitudes of HLC formation depending on meteorological conditions. The results of lidar measurements were aligned with the ERA5 reanalysis data [22] at each point of the vertical profile from 0 to 15 km with a step of 150 m. Vertical profiles of the following meteorological parameters were generated for each session, taking into account the ERA5 data: relative humidity, absolute humidity, wind speed, wind direction, and temperature [18]. After creating the database and bringing the dataset to a format that allows for the training of algorithms, we created a neural network to solve the following atmospheric optics problems:

(1) Determine the probability of observation of HLCs depending on meteorological parameters (classification task);
(2) Make a preliminary estimation of the observation altitude and boundaries of HLCs depending on meteorological parameters (regression task);
(3) Estimate BSPM values using meteorological parameters (regression task).

We encountered a problem; despite the substantial number of lidar experiments, the number of measurement series in the 2016–2023 period, during which HLCs were recorded and which are suitable for training the neural network, was only 312. The problem of statistical smallness of the experimental dataset was solved by software, which is acceptable at the step of software prototype development. The next step requires more data to refine the result and to add more information about the environment in the area of the experiment.

The augmentation of the experimental dataset with data is ongoing. In addition, the information about the environment in the experimental area is being investigated, the characteristics of which will be added as input parameters for training the neural network.

3. Results

Knowing the values of vertical profiles of meteorological parameters, we investigated their relationship with BSPM. In turn, neural networks allow us to reconstruct almost any dependence given a sufficient amount of data. It was stated above that there were 312 cases of HLC observations, which is rather small for such a task. This fact imposes a restriction on the maximum size of the neural network and the number of parameters to be trained, which should not exceed the amount of data.

The meteorological data correlation study was divided into the following steps:

- Dimensionality reduction in the ERA5 reanalysis data;
- Analysis of the relationship between the altitude of HLC detection and meteorological parameters;
- Determination of BSPM based on meteorological parameters.

The presence of the first step is motivated by the fact that if we use the values from the vertical profiles as is, it will lead to a significant increase in the number of neural network parameters.

In addition, we have significantly expanded the database by conducting many atmospheric sensing experiments in recent years. New algorithms have been created to data process and compare them with the meteorological situation.

3.1. Preliminary Analysis

As a first step in the analysis, the change in BSPM elements with altitude within the HLC thickness was investigated. Figure 2 shows an example of one BSPM element measurement made on 19 May 2016, starting at 14:07. The m_{11} element is always equal to one because all BSPM elements are normalized to it. The remaining plots show the behavior of the values of all BSPM elements within the HLC altitude range. The element values in both channels of the lidar receiving system are consistent with each other and correspond to some constant value with some noise added. Similar behavior is present in the other dimensions. Thus, altitude dependence within the HLC is not observed, and for one observation, it is sufficient to take the median value of the BSPM element for the channel with the best signal-to-noise ratio, which is more resistant to the presence of outliers.

The next step was to study the distributions of values of the BSPM elements as functions of meteorological parameters. Figure 3 presents histograms of median values of BSPM elements in the range of HLC altitudes. Since the measurements were performed under different meteorological conditions and for different altitudes of HLC formation, we can assume that we should obtain histograms "smeared" in some interval, since their construction did not consider the presence of dependencies on meteorological parameters. Figure 3 shows that this is fulfilled only for the following BSPM elements: m_{22}, m_{33}, and m_{44}. For the remaining elements, no variability is observed. For elements m_{12}, m_{21}, m_{13}, m_{31}, m_{14}, and m_{41}, a distribution like a normal distribution is observed, with a mean of 0 and some small variance at the noise level. The background lidar signal value is used as a noise. This value is calculated as the average lidar signal for the upper 3 km of the lidar operation altitude range (12–15 km) for each receiving channel. For elements m_{24}, m_{42}, m_{34}, and m_{43}, the similarity to a normal distribution is no longer observed, but the mean is also observed in the zero region. From this, we can conclude that the most sensitive to environmental conditions are the following BSPM elements: m_{22}, m_{33}, and m_{44}.

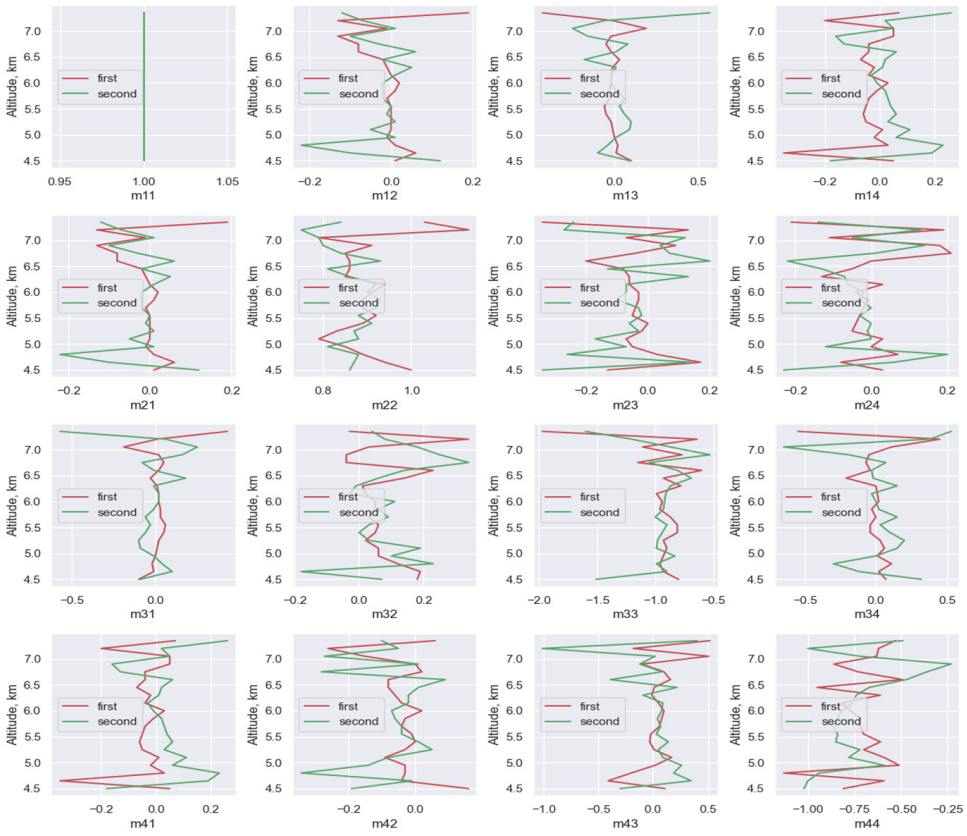

Figure 2. An example of the BSPM values of the elements' dependence on the altitude within the HLC thickness obtained in measurements with the first (parallel) and second (parallel) channels of the lidar receiving system, respectively.

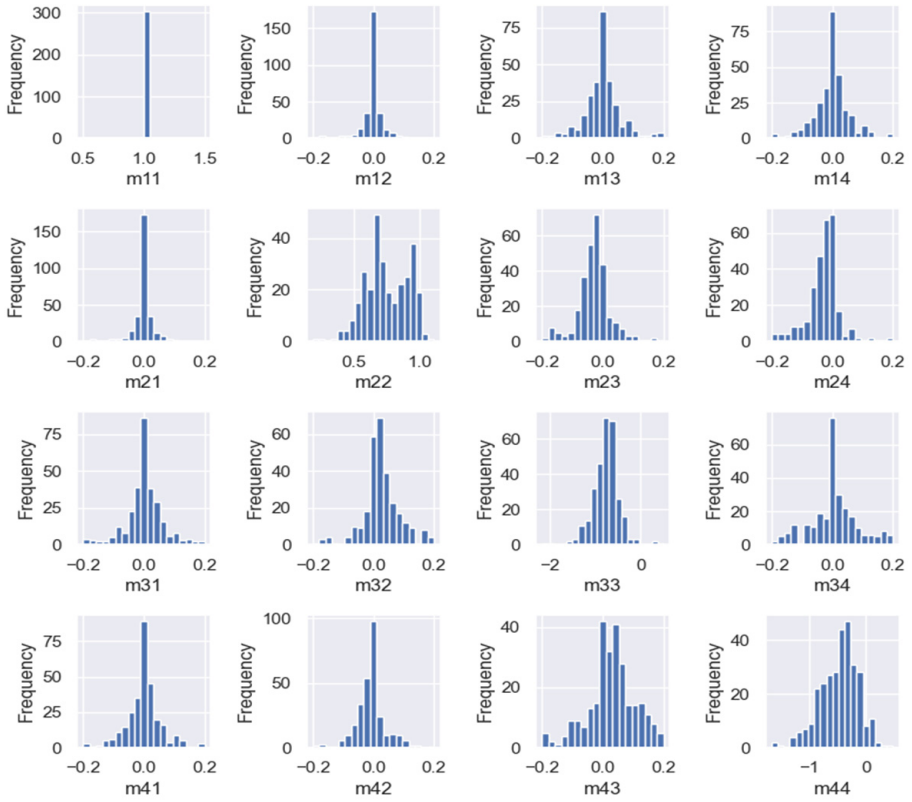

Figure 3. Distribution of BSPM element values.

The last step is to make sure that there is some dependence between the meteorological parameters and the values of the HLC BSPM elements. To investigate this question, instead of the altitude profile, the value of the meteorological parameter in the central height region of the HLC was taken. Figures 4–6 show the scatter plot for the m_{22}, m_{33}, and m_{44} BSPM elements. In pressure and temperature diagrams, there is no distribution of elongated points along one of the axes, which indicates that there is some dependence between the values. This was implemented as a proof of concept to test for any dependency on the experimental environment. So, if we obtain any shape that differs from the ellipse/band, then we have some dependency. And we obtained a non-ellipse/non-band shape for some coefficients. Therefore, it makes sense to try machine learning to restore this dependency. Further, in the article, we provide an analysis of the altitude profiles of meteorological parameters. The center point was taken only for simplification of view.

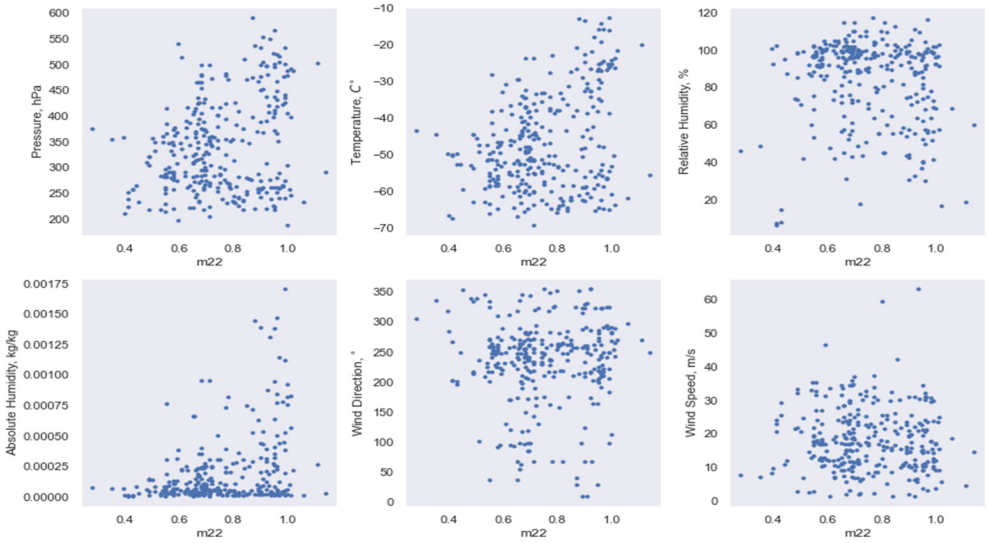

Figure 4. Atmosphere meteorological characteristics scatter plot depending on m_{22} BSPM element.

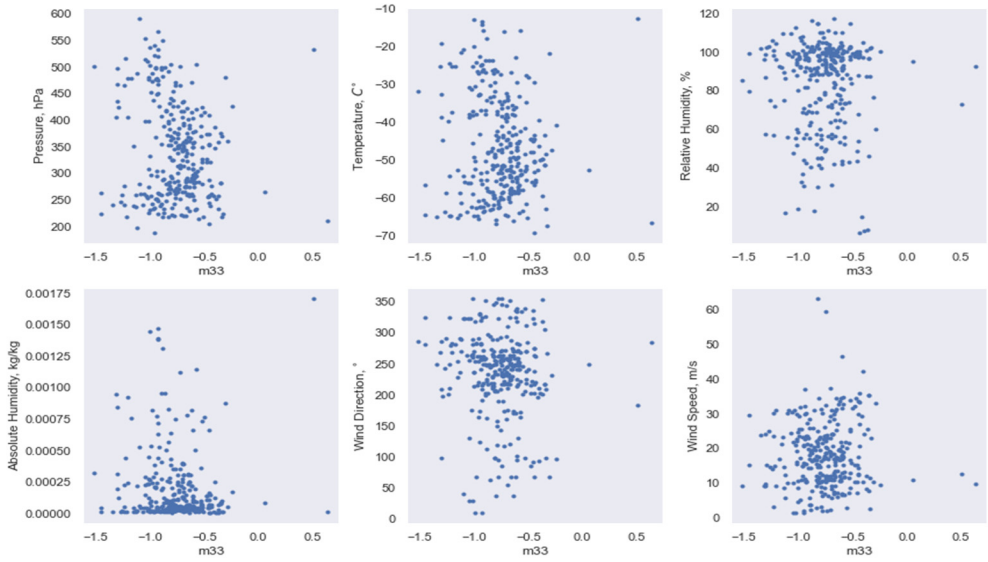

Figure 5. Atmosphere meteorological characteristics scatter plot depending on m_{33} BSPM element.

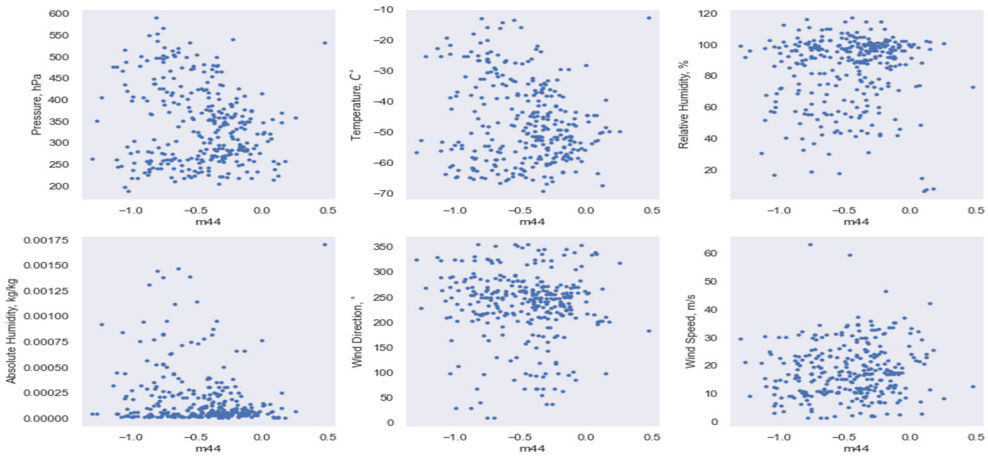

Figure 6. Atmosphere meteorological characteristics scatter plot depending on m_{44} BSPM element.

Figures 7–10 show the scatter plots for the m_{24}, m_{42}, m_{34}, and m_{43} BSPM elements. In all figures, there is a well-defined vertical trend, which signals weak or no dependence between the values.

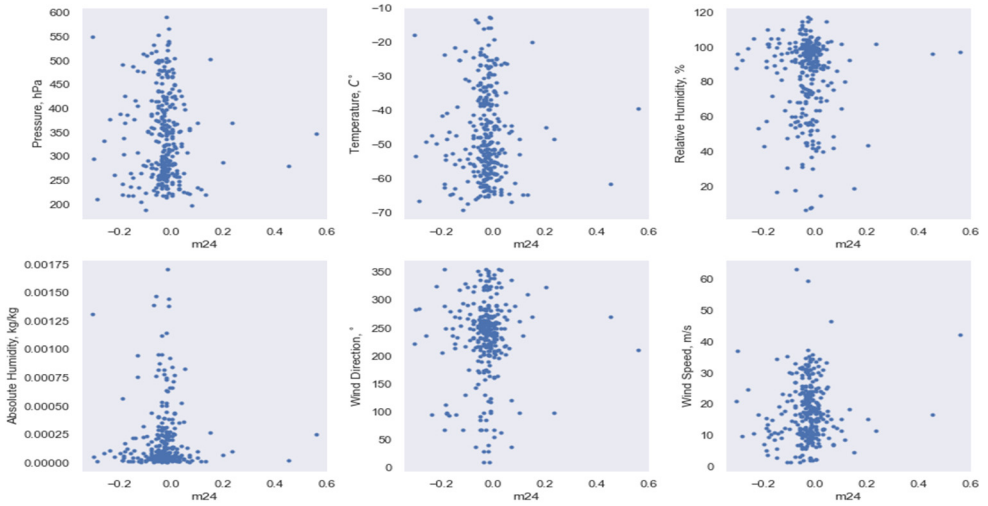

Figure 7. Atmosphere meteorological characteristics scatter plot depending on m_{24} BSPM element.

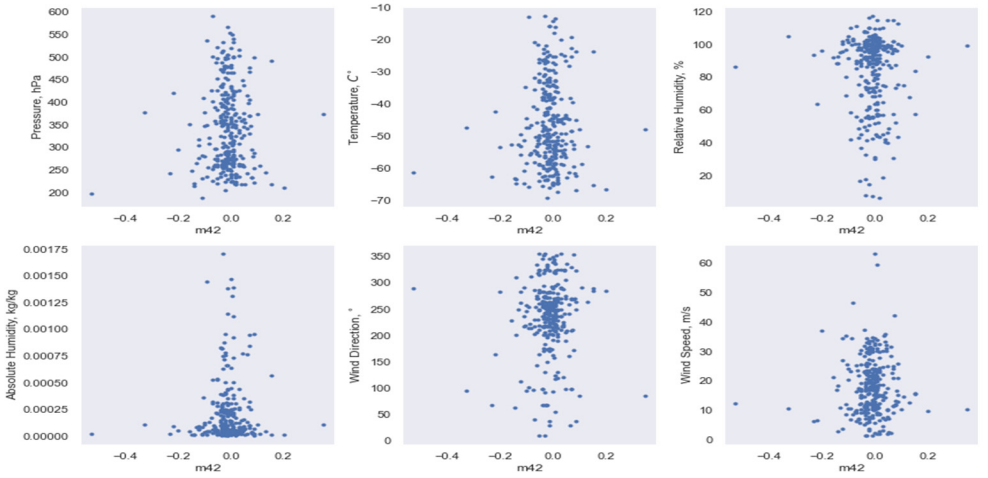

Figure 8. Atmosphere meteorological characteristics scatter plot depending on m_{42} BSPM element.

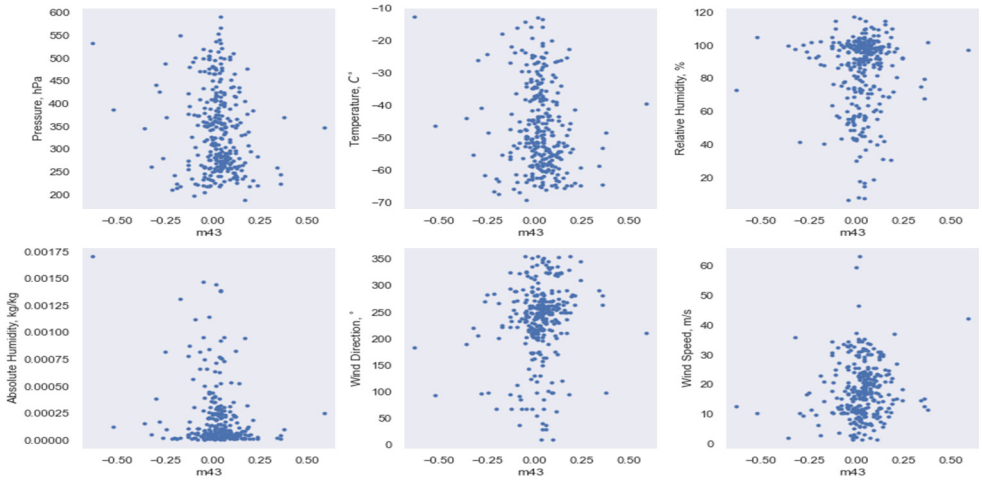

Figure 9. Atmosphere meteorological characteristics scatter plot depending on m_{43} BSPM element.

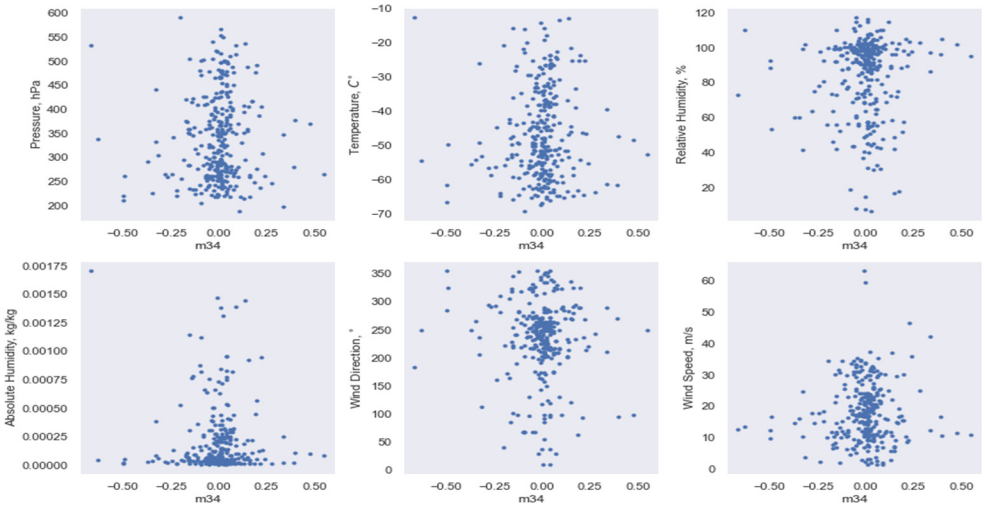

Figure 10. Atmosphere meteorological characteristics scatter plot depending on m_{34} BSPM element.

Thus, the following HLC BSPM elements are subject to analysis using machine learning methods: m_{22}, m_{33}, and m_{44}. This is probably due to the fact that the element m_{44} and the sum m_{22} + m_{33} are invariant with respect to the rotation of the lidar basis (or the cloud itself) relative to the vertical axis [23]. The rest of the elements do not reveal dependences with meteorological parameters or require much more data.

3.2. Implementation of Data Dimensionality Reduction

A total of 124,512 altitude profiles of meteorological parameters with hourly resolution were obtained using the ERA5 reanalysis for the period from 2009 to 2023. Each profile corresponds to the lidar coordinates and consists of more than 30 points on a heterogeneous pressure grid. For standardization, all profiles were transformed by interpolation to a single 31-point elevation grid. Feeding all of these values to the input of a neural network will result in the number of parameters of this network increasing significantly, which creates the need for data compression with maximum information preservation. The classical method of dimensionality reduction is principal component analysis (PCA). This method uses a singular value decomposition of the covariance matrix of the data. The spectrum of this decomposition characterizes the components that carry the most information.

An alternative approach is the use of so-called autoencoders, which are special neural networks. Such a tool can be represented as an ordinary multilayer perceptron. Its specific feature is that during training, they are required to obtain the same values at the output as at the input. In this case, each layer of the neural network is an encoder that translates the

vector of input values into a new linear space of a different dimension. If there is a layer with a small number of neurons inside the autoencoder, it leads to data compression. The outputs of this layer can be used as new components containing maximum information from the point of view of this autoencoder. In some cases, such neural networks obtain better data compression than PCA, as they take into account nonlinear dependencies between input data. In addition, the choice of the activation function allows for customization of the mapping scale of the input data to the compressed representation. We will use the hyperbolic tangent as such a function: this will allow us to obtain components in the range from –1 to 1. The activation function of the output layer will be linear, which will allow us to obtain arbitrary values at the output of the neural network. In addition, we need to take into account the fact that the input data have different scales of change. This complicates the use of a hyperbolic tangent, which may contribute to the frozen state of some neurons. This in turn results in values that are too large or too small, forcing the neuron to keep its state at –1 or 1, which disturbs the training of the network. To eliminate this problem, all input data are normalized by subtracting the mean from them and dividing by the standard deviation. This produces input values with a mean of 0 and a standard deviation of 1.

To determine the dimensionality of the inner compressive layer of the autoencoder, as well as the constraint on the singular spectrum in PCA, we considered the standard deviation between the original altitude profiles and those reconstructed from the compressed components, determined according to the following formula:

$$MSE = \sqrt{\frac{1}{N}\sum\left(y_{predict} - y_{data}\right)^2} \quad (1)$$

If increasing the number of compressed components did not significantly improve this metric, the increase in components was stopped. During training, the data were divided randomly into two samples: training and test. The test sample represents 33% of the total data. Table 1 shows the standard deviation values for PCA and autoencoder (AE) obtained on the test sample. In most cases, AE gives better compression than PCA. The exception is absolute humidity, for which these approaches give comparable results.

Table 1. RMS deviation of meteorological parameters reconstructed from compressed vertical profile data.

	Temperature (°C)	**Relative Humidity (kg*kg^{-1})**	**Absolute Humidity (%)**	**Wind Speed (m/s^2)**
PCA	2.36	60.7	0.57×0^{-7}	1.61
AE	1.66	42.54	0.51×10^{-7}	1.58

The following values for the number of compressed components were found during training: temperature profile requires three components; relative humidity profile requires six components; absolute humidity profile requires three components; and wind speed profile requires five components.

3.3. Estimation of HLC Detection Altitude

One of the important characteristics of HLCs, in addition to the BSPM elements, are the altitudes of their lower and upper boundaries. It is of interest to determine whether there is a relationship between meteorological parameters and HLC detection altitudes. For this purpose, it was decided to move from the determination of the upper and lower boundary to the determination of the altitude of the cloud center and the deviation of the boundaries from this center. To counterbalance the scale of the center altitude variation, normalization was performed: 8 km was subtracted and divided by two (average HLC altitude is about 6–10 km, thickness—4 km. We take the center of this interval and divide by half of the thickness). This allows us to further obtain a more stable process of neural network training. The normalization parameters were taken from the general distribution of altitudes. Due to

the fact that there are data from lidar experiments from 2009, in which the HLC altitudes were also determined, the set of available values increases to 779 observations.

To determine the quality of neural network performance, we used the cross-validation approach, which is convenient to apply in conditions with a small amount of data. In this approach, the data are randomly divided into K equal parts, so-called folds. Then, the same steps are performed for each part:

- The current part forms the test sample;
- The remaining parts form the training sample;
- Training of the neural network on the training sample and calculation of the standard deviation on the test sample.

As a result, we obtain K different values of standard deviations and K trained neural networks. In the case of normal training and the absence of data heterogeneity, these values will be commensurable in values. Otherwise, we will obtain quite different values.

In addition, the random forest (RF) method with the number of trees equal to 100 was used as a reference. This method is convenient because it is robust to different scales of input data changes and is not prone to overtraining. Thus, if the neural network obtains a result worse than the random forest method, it becomes a sign that the network architecture is chosen incorrectly. It is also a benefit that the random forest method allows us to identify those input parameters that give the greatest contribution to the determination of the output value, which can also provide additional information for analysis. Thus, it is interesting to study the behavior of the random forest method on full data and on compressed ones. At the same time, it is important to understand that, in some cases, the set of parameters obtained in this way may lead to misinterpretation of the data. This information can only be used for auxiliary purposes.

In addition, point diagrams are a convenient study tool to evaluate the relationship of one value to another. Figure 11 shows the scatter plots for each fold using the random forest method and compression with PCA.

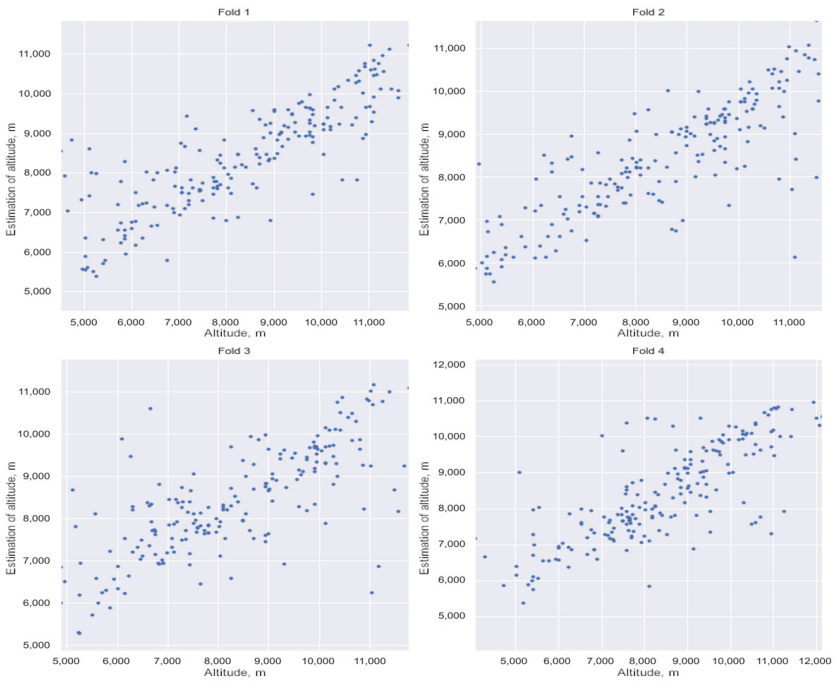

Figure 11. Scatter plot of HLC center altitude determination using random forest and PCA compression method.

In the case of perfect altitude determination, we should see a straight line as the estimate will coincide with the true value. Due to the presence of noise or weak dependence on the input data, the estimate will differ from the true value. In the figure, we can see

a linear relationship between the estimate and the true value. In addition, there are point deviations that give a very large error. Presumably, this is due to either an error in the experimental determination of the HLC altitude or to the specific circumstances of its formation.

It is convenient to consider the estimation error in terms of standard deviation. The values of this deviation can be seen in Table 2 (first column). The magnitude of the error is of the order of 1 km but with low variation, signaling the homogeneity of the data and the absence of specific outliers.

Table 2. The value of standard deviation in HLC altitude estimation (in meters) for random forest and neural network method with different data compression methods.

	RF (PCA)	NN (PCA)	RF (AE)	NN (AE)
Fold 1	1097.07	1232.56	1162.26	1320.61
Fold 2	1048.35	1350.06	1065.12	1371.62
Fold 3	1177.63	1433.94	1181.13	1383.15
Fold 4	1095.24	1222.92	1150.78	1388.10

For the neural network, we have chosen a model of an ordinary multilayer perceptron with one hidden layer of 15 neurons and an activation function in the form of a hyperbolic tangent. The output neuron has a linear activation function since we are trying to solve a regression problem. This architecture corresponds to 316 training parameters, which is smaller than the training dataset. Thus, it will be more difficult for the neural network to be overtrained. Figure 12 shows the scatter plots obtained with the neural network and data compression using PCA. A linear relationship can also be seen here. The values of the standard deviation are presented in Table 2 (second column). The values are of the same order of magnitude as the random forest method but exceed it. This is most likely due to insufficient data, making the training of a large network vulnerable to overtraining and a small network insufficient to reconstruct the dependency. Increasing the number of neurons in the hidden layer leads to a rapid overtraining of the network and an increase in the error, while decreasing it leads to an inability to reconstruct the dependency and also to an increase in the error.

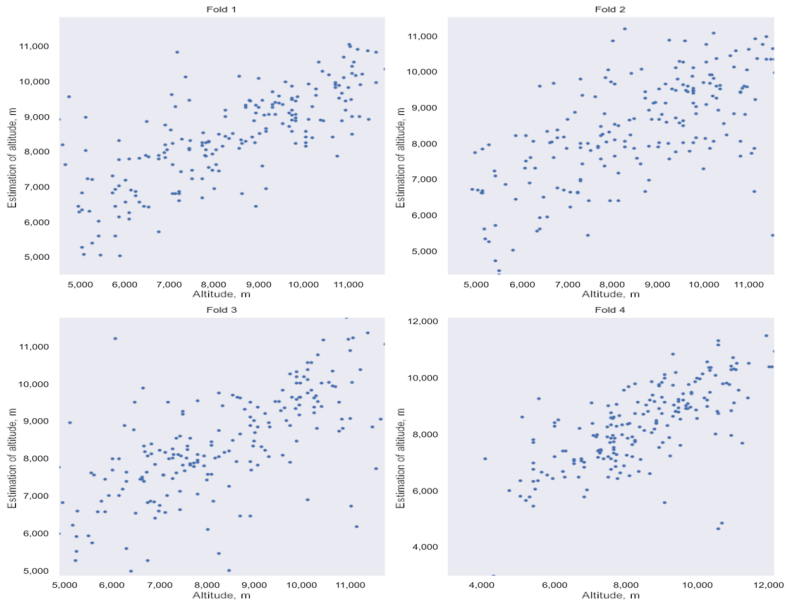

Figure 12. Scatter plot of HLC center altitude determination using neural network and PCA compression.

Figures 13 and 14 show scatter plots using autoencoder compression. The error rates are comparable to PCA compression. The use of compressed data also allows for the estimation of observed values, and both methods give comparable results.

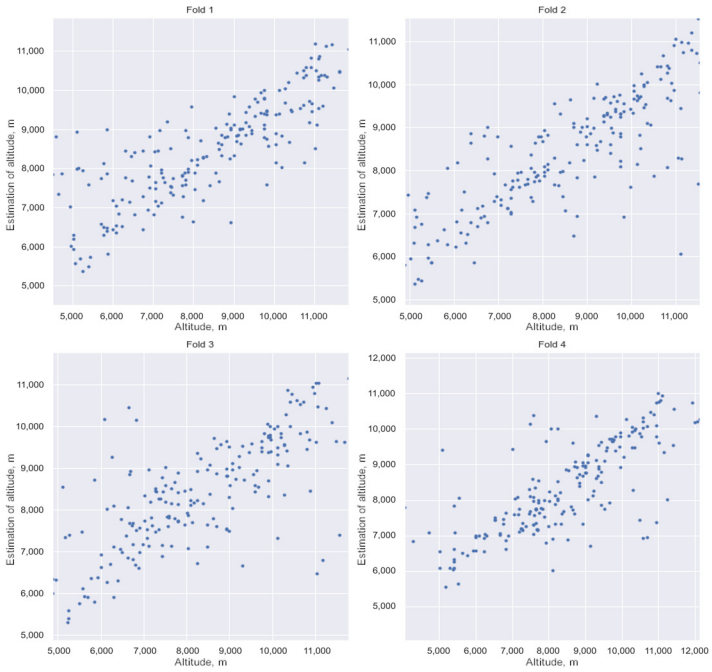

Figure 13. Scatter plot of HLC center altitude determination using the random forest and AE compression method.

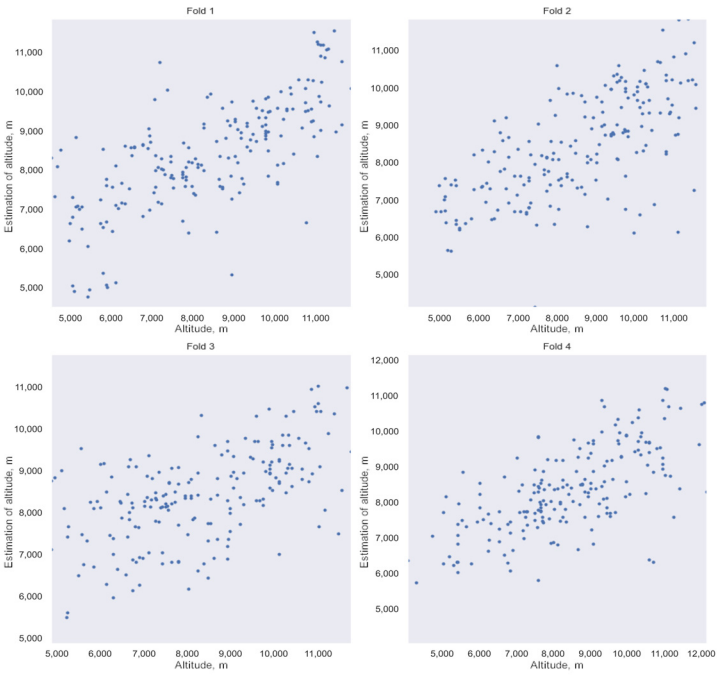

Figure 14. Scatter plot of HLC center altitude determination using neural network and AE compression.

In all cases, there is a tendency to be able to estimate altitude using machine learning methods and meteorological observations. However, the magnitude of the error is too high to use these results for practical applications. This can be corrected either by adding additional specific data, such as taking into account the dynamics of profile changes over

time, or taking into account anthropogenic factors. It is also necessary to expand the experimental dataset, which will allow us to obtain more unambiguous results.

3.4. Defining HLC Boundaries

Similarly to Section 3.3, the same calculations were performed, but this time, it was to estimate the displacement of the boundary location relative to the center of the cloud. The displacement itself was divided by two to bring it to a scale convenient for working with neural networks. Table 3 summarizes the results of the calculations.

Table 3. The value of standard deviation in estimating the off-center displacement of HLC boundaries (in meters) for the random forest method and neural network with different data compression methods.

	RF (PCA)	NN (PCA)	RF (AE)	NN (AE)
Fold 1	458.60	557.44	458.91	548.99
Fold 2	420.30	494.86	454.43	510.57
Fold 3	491.13	520.78	498.08	582.19
Fold 4	488.48	603.55	488.23	549.19

Figures 15 and 16 show plots of the boundary off-center displacements of the HLC with PCA compression. It can be seen from them that there is some correspondence between the displacement estimate and the displacement measured experimentally.

It is worth noting that for the neural network, this dependence is poorly observed. Thus, it can also be noted here that there is a relationship between meteorological parameters, but to increase the quality of the assessment, additional information and expansion of the experimental dataset are required.

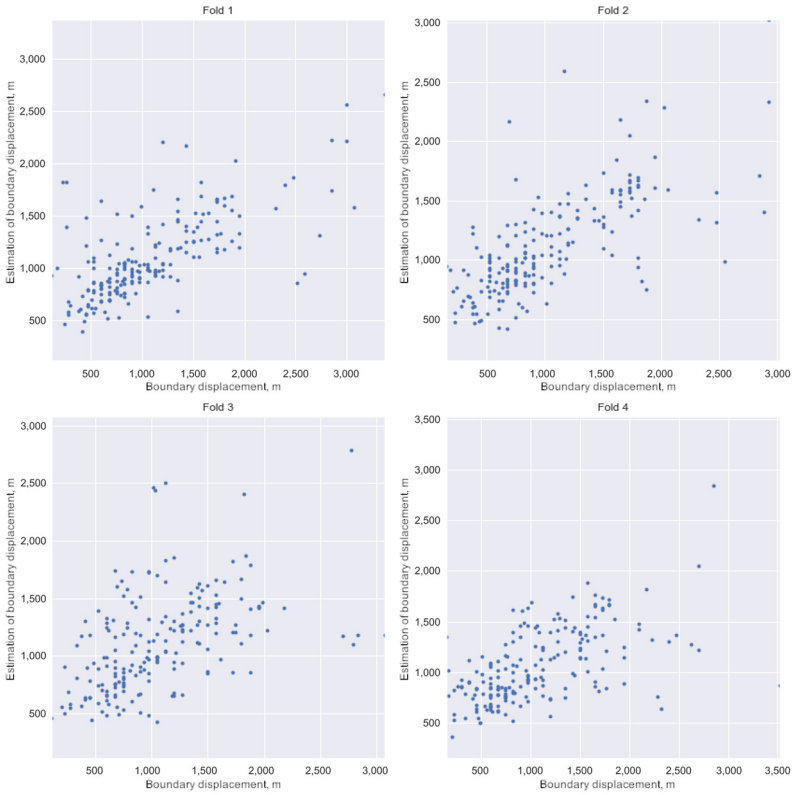

Figure 15. Scatter plot of HLC boundary displacement determination using the random forest and PCA compression method.

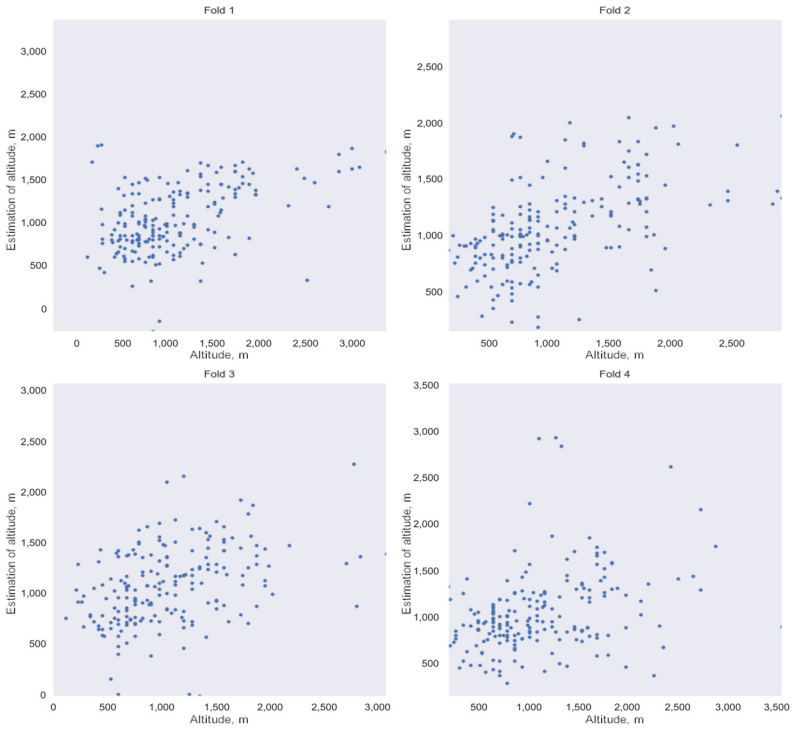

Figure 16. Scatter plot of HLC boundary displacement determination using neural network and PCA compression.

3.5. Evaluation of HLC BSPM Elements

In this paper, the elements m_{22}, m_{33}, and m_{44} of the HLC BSPM were evaluated. There were 312 measurements suitable for training, which significantly complicated the task due to the small size of the experimental array. A random forest method with 100 trees was used as a reference. This allowed us to estimate the potentially optimal result. As a neural network, we took a multilayer perceptron with a hidden layer of five neurons—the simplest model for small statistics. To evaluate the quality of these approaches, we also used the value of standard deviation and cross-validation of two folds. The following results were obtained.

Table 4 presents data for the m_{22} BSPM element. The worst result is observed when using the neural network method, while the random forest method yields better results.

Table 4. The value of standard deviation in estimating the m_{22} BSPM element for random forest method and neural network with different data compression methods.

	RF (PCA)	NN (PCA)	RF (AE)	NN (AE)
Fold 1	0.12	0.14	0.12	0.17
Fold 2	0.14	0.16	0.13	0.19

The scatter plot in Figure 17 shows a weak relationship between the estimate of m_{22} and its measured value. This is most likely due to the weak relationship with the input parameters. In Figure 18, it can be seen that the neural network is almost unable to detect the relationship between the compressed meteorological parameters and the element m_{22}.

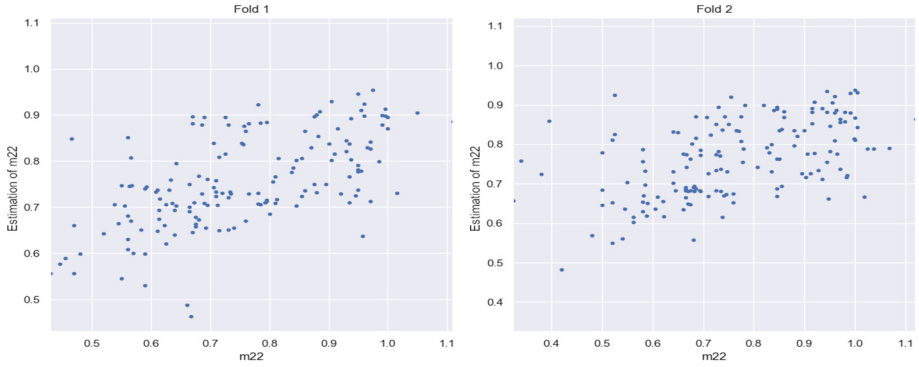

Figure 17. Scatter plot of m_{22} estimation using the random forest and PCA compression method.

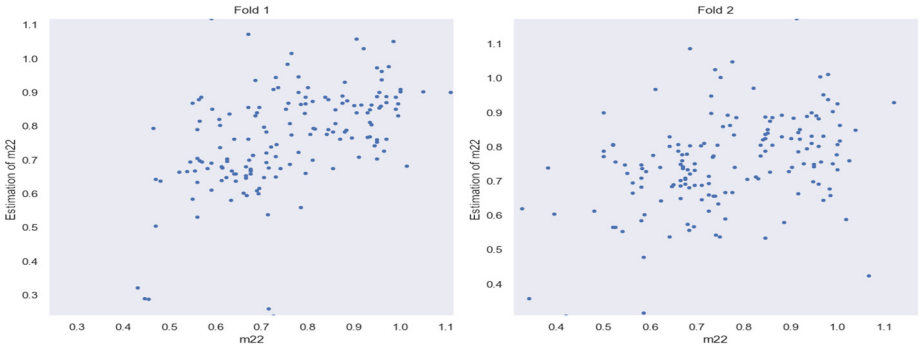

Figure 18. Scatter plot of m_{22} BSPM estimation using neural network and PCA compression.

Table 5 shows the m_{33} BSPM element. The best result is observed when using the random forest method. Both methods virtually did not determine the presence of dependencies (Figures 19 and 20), i.e., this BSPM element is independent of meteorological conditions.

Table 5. The value of standard deviation in estimating the m_{33} BSPM element for random forest method and neural network with different data compression methods.

	RF (PCA)	NN (PCA)	RF (AE)	NN (AE)
Fold 1	0.17	0.20	0.18	0.21
Fold 2	0.21	0.25	0.22	0.22

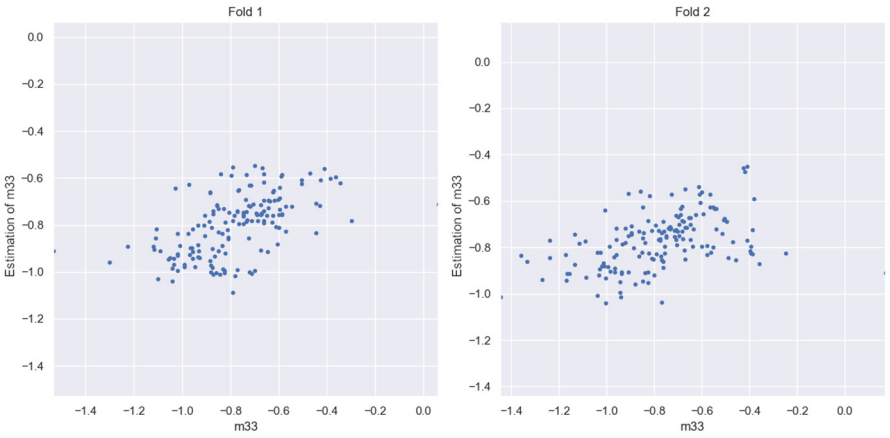

Figure 19. Scatter plot of m_{33} BSPM estimation using the random forest and PCA compression method.

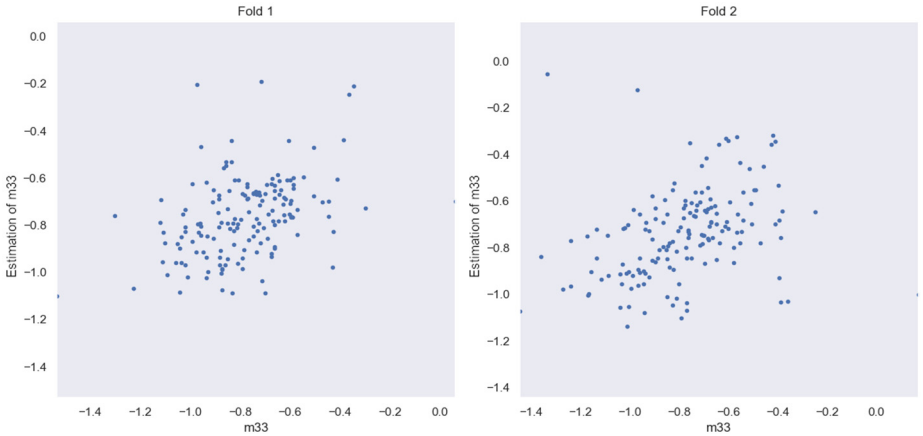

Figure 20. Scatter plot of m_{33} BSPM estimation using neural network and PCA compression.

Table 6 shows the m_{44} BSPM element. A better result is observed for the random forest method. In both cases, only a small dependence on the input values is observed (Figures 21 and 22).

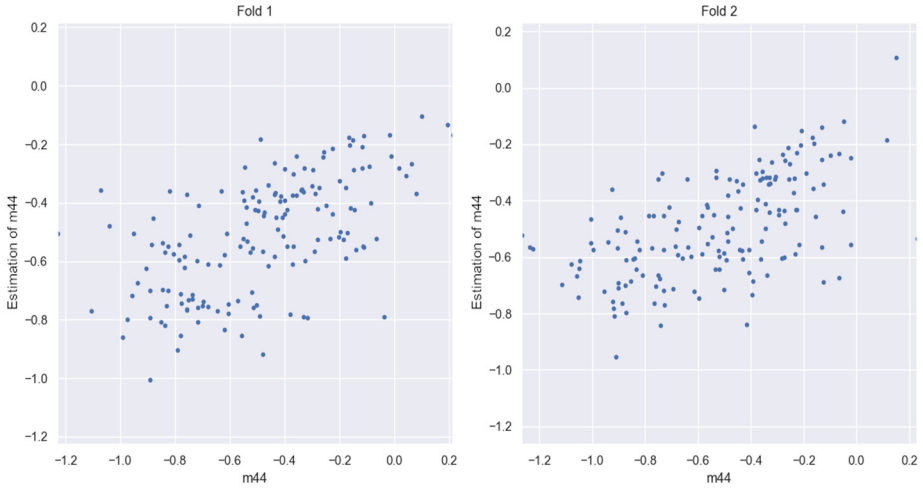

Figure 21. Scatter plot of m_{44} BSPM determination using the random forest and PCA compression method.

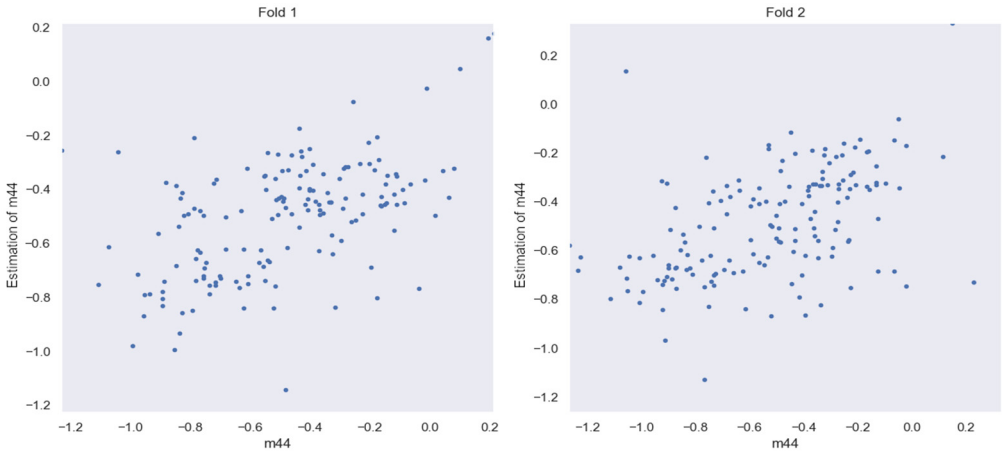

Figure 22. Scatter plot of m_{44} BSPM estimation using neural network and PCA compression.

Table 6. The value of standard deviation in estimating the m_{44} BSPM element for random forest method and neural network with different data compression methods.

	RF (PCA)	NN (PCA)	RF (AE)	NN(AE)
Fold 1	0.24	0.25	0.18	0.20
Fold 2	0.25	0.28	0.21	0.22

Concluding the description of the results, it is worth recalling that we considered the elements of the HLC BSPM to be normalized by m_{11}; because of this, m_{11} in the analyzed matrices is equal to one. The elements m_{22} and m_{44} of such BSPMs depend on meteorological parameters. Element m_{33} requires further study and expansion of the analyzed dataset. The enrichment of the experimental dataset is continued [18,21]. In addition, the influence of other atmospheric parameters on the HLC BSPM elements is being investigated. These parameters will be added as input data for neural network training in the future.

4. Discussion

The software prototype based on an artificial neural network is a software application that uses machine learning algorithms to analyze meteorological observation data and predict the optical and geometric characteristics of HLC. The prototype can be used to test different models and determine the most effective approach. It can also be used to demonstrate the capabilities of the final product. The resulting tool allows a preliminary evaluation of the BSPM elements, boundaries, and detection altitudes of the HLC. At the moment, a weak dependence on the meteorological parameters for some HLC parameters can be noted. But to answer this question unambiguously, more data are needed to clarify the result and to add additional information about the environment in the area of the experiment, which we plan to realize in the near future.

5. Conclusions

Analysis of the obtained results showed that there is a dependence on meteorological parameters for the following values: the altitude of the HLC center; elements m_{22} and m_{44} of BSPM (with m_{11} always equal to 1); and the m_{33} element of BSPM requires further study and expansion of the analyzed dataset. For the rest of the values, a sufficiently large error value is observed, and at the moment, it is impossible to give an unambiguous answer. At the same time, the result is not affected by the choice of data compression method to reduce their dimensionality.

In almost all cases, the random forest method gave better results than the simple multilayer perceptron. This is most likely caused by two factors: preprocessing of input data and a small amount of experimental data suitable for training. As for preprocessing, the usual normalization (zero mean and variance 1) was chosen, which may not be sufficient. As further steps, it is necessary to consider transformations that bring the histogram distribution of the input data to a symmetric form and more similar to a normal distribution.

The size of the data also affected the performance of the neural network. As the neural network becomes more complex, it is able to approximate more sophisticated functions, but a larger amount of data must be used. Using small amounts of data leads to overtraining of the network and hence deterioration in the results obtained. To improve the results and to establish additional possible dependencies between meteorological parameters and HLC characteristics, we plan to conduct research in the preprocessing of meteorological parameters, namely taking into account the dynamics of change over time, the gradient in the area of the experiment, and the anthropogenic factor. It is also necessary to increase the volume of the experimental dataset (performing new experiments on laser sensing of the atmosphere).

Author Contributions: O.K., M.P., I.B. (Iurii Bordulev) and V.K. developed a prototype software product based on artificial neural networks for meteorological observation data analysis and prediction of optical and geometric characteristics of HLC. They also analyzed the results obtained and partially investigated the applicability feature of neural networks to small datasets. In addition, these participants analyzed the effect of data preprocessing on the efficiency of neural network training. I.B. (Ilia Bryukhanov), E.N., A.D., I.Z. and I.S. developed the methodology of the experiments, built the HAMPL of NR TSU, and performed lidar measurements. I.B. (Ilia Bryukhanov) and E.N. processed lidar data. I.B. (Ilia Bryukhanov) together with A.D. processed the meteorological data. I.S. supervised the experiments on the HAMPL of NR TSU and the processing of lidar and meteorological data. O.K., M.P., I.B. (Iurii Bordulev), V.K., I.B. (Ilia Bryukhanov), E.N., A.D., I.Z., S.V. and I.S. discussed the results obtained and wrote and edited the text of this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Russian Science Foundation, Grant № 21-72-10089.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Yeganeh-Bakhtiary, A.; EyvazOghli, H.; Shabakhty, N.; Kamranzad, B.; Abolfathi, S. Machine learning as a downscaling approach for prediction of wind characteristics under future climate change scenarios. *Complexity* **2022**, *2022*, 8451812.
2. Donnelly, J.; Abolfathi, S.; Pearson, J.; Chatrabgoun, O.; Daneshkhah, A. Gaussian process emulation of spatio-temporal outputs of a 2D inland flood model. *Water Res.* **2022**, *225*, 119100. [CrossRef] [PubMed]
3. Nourani, V.; Khodkar, K.; Paknezhad, N.J.; Laux, P. Deep learning-based uncertainty quantification of groundwater level predictions. Stoch. Environ. *Res. Risk Assess* **2022**, *36*, 3081–3107. [CrossRef]
4. Zennaro, F.; Furlan, E.; Simeoni, C.; Torresan, S.; Aslan, S.; Critto, A.; Marcomini, A. Exploring machine learning potential for climate change risk assessment. *Earth-Sci. Rev.* **2021**, *220*, 103752. [CrossRef]
5. Sattari, M.T.; Mirabbasi, R.; Sushab, R.S.; Abraham, J. Prediction of groundwater level in Ardebil plain using support vector regression and M5 tree model. *Groundwater* **2018**, *56*, 636–646. [CrossRef] [PubMed]
6. Kim, I.; Kim, B.; Sidorov, D. Machine Learning for Energy Systems Optimization. *Energies* **2022**, *15*, 4116. [CrossRef]
7. Feigelson, E.M. (Ed.) *Radiation Properties of Perispheric Clouds*; Nauka: Moscow, Russia, 1989. (In Russian)
8. Winker, D.M.; Trepte, C.R. Laminar cirrus observed near the tropical tropopause by LITE. *Geophys. Res. Lett.* **1998**, *25*, 3351–3354. [CrossRef]
9. Liou, K.N. Influence of cirrus clouds on weather and climate processes: A global perspective. *J. Geophys. Res.* **1986**, *103*, 1799–1805. [CrossRef]
10. Sassen, K.; Griffin, M.K.; Dodd, G.C. Optical scattering, and microphysical properties of subvisual cirrus clouds, and climatic implications. *J. Appl. Meteorol.* **1989**, *28*, 91–98. [CrossRef]
11. Dmitrieva-Arrago, L.R.; Trubina, M.A.; Tolstyh, M.A. Role of Phase Composition of Clouds in Forming High and Low Frequency Radiaion. *Proc. Hydrometeorol. Res. Cent. Russ. Fed.* **2017**, *363*, 19–34. (In Russian)
12. Stengel, M.; Meirink, J.F.; Eliasson, S. On the Temperature Dependence of the Cloud Ice Particle Effective Radius—A Satellite Perspective. *Geophys. Res. Lett.* **2023**, *50*, e2022GL102521. [CrossRef]
13. Scientific and Technological Infrastructure of the Russian Federation. Radiophysical Complex: High-Altitude Polarization Lidar for Atmospheric Sensing and Tomsk Ionospheric Station "LIDAR-IONOSONDE". Available online: https://ckp-rf.ru/catalog/usu/73573 (accessed on 25 November 2023).
14. Guasta, M.D.; Vallar, E.; Riviere, O.; Castagnoli, F.; Venturi, V.; Morandi, M. Use of polarimetric lidar for the study of oriented ice plates in clouds. *Appl. Opt.* **2006**, *45*, 4878–4887. [CrossRef] [PubMed]
15. Hayman, M.; Thayer, J.P. General description of polarization in lidar using Stokes vectors and polar decomposition of Mueller matrices. *J. Opt. Soc. Am.* **2012**, *29*, 400–409. [CrossRef] [PubMed]
16. Volkov, S.N.; Samokhvalov, I.V.; Cheong, D.H.; Kim, D. Investigation of East Asian clouds with polarization light detection and ranging. *Appl. Opt.* **2015**, *54*, 3095–3105. [CrossRef]
17. Kokhanenko, G.P.; Balin, Y.S.; Klemasheva, M.G.; Nasonov, S.V.; Novoselov, M.M.; Penner, I.E.; Samoilova, S.V. Scanning polarization lidar LOSA-M3: Opportunity for research of crystalline particle orientation in the ice clouds. *Atmos. Meas. Tech.* **2020**, *13*, 1113–1127. [CrossRef]
18. Kuchinskaia, O.; Bryukhanov, I.; Penzin, M.; Ni, E.; Doroshkevich, A.; Kostyukhin, V.; Samokhvalov, I.; Pustovalov, K.; Bordulev, I.; Bryukhanova, V.; et al. ERA5 Reanalysis for the Data Interpretation on Polarization Laser Sensing of High-Level Clouds. *Remote Sens.* **2023**, *15*, 109. [CrossRef]

19. Central Aerological Observatory. Available online: http://cao-ntcr.mipt.ru/monitor/locator.htm (accessed on 1 February 2024).
20. University of Wyoming. Available online: http://weather.uwyo.edu (accessed on 1 February 2024).
21. Penzin, M.S.; Bryukhanov, I.D.; Kuchinskaia, O.I.; Ni, E.V.; Pustovalov, K.N.; Zhivotenyuk, I.V.; Doroshkevich, A.A.; Bordulev Iu, S.; Samohvalov, I.V. Verification of ERA5 reanalysis data for the interpretation of lidar investigation of high-level clouds. In Proceedings of the SPIE 28th International Symposium on Atmospheric and Ocean Optics: Atmospheric Physics, Tomsk, Russia, 4–8 July 2022; Volume 12341, p. 4. [CrossRef]
22. Copernicus Climate Data Store. Available online: https://cds.climate.copernicus.eu (accessed on 1 February 2024).
23. Kaul, B.V.; Samokhvalov, I.V.; Volkov, S.N. Investigating particle orientation in cirrus clouds by measuring backscattering phase matrices with lidar. *Appl. Opt.* **2004**, *43*, 6620–6628. [CrossRef]