Prediction of high-level cloud parameters based on polarization lidar observations using machine learning tools

M. Penzin · O. Kuchinskaia · I. Akimov · D. Romanov · I. Bryukhanov · I. Samokhvalov

Received: 4 September 2024 / Accepted: 4 October 2024

Abstract
The paper focuses on machine learning methods in atmospheric laser remote sensing. Atmospheric lidar data have been stored since 2009, when high-altitude matrix polarization lidar measurements became systematic. Light detection and ranging (lidar) data are used to determine optical (backscattering phase matrix (BSPM), thickness, and scattering ratio) and geometrical (altitude of lower and upper boundary, vertical extent) parameters of higher-level clouds. The data storage is then added by meteorological parameters derived from radiosonde observations and ERA5 reanalysis data (temperature, relative and specific humidity, wind speed and direction). Methodology includes machine learning models, random forest, and modified version incorporating principal component analysis for dimensionality reduction to probe the complex relationship between meteorological parameters and specific BSPM elements. For the first time, different lidar measurement techniques identify two distinct maxima in the distribution of BSPM elements.

Keywords Atmosphere · High-level clouds · Machine learning · Random forest · Principal component analysis · Polarization lidar · Backscattering phase matrix · Radiosonde observation · ERA5 reanalysis

Introduction

Increasingly noticeable climate changes necessitate the improvement of weather forecasts. The accuracy improvement of forecasting the atmospheric conditions, requires completeness, higher spatial and temporal resolution, more exact weather forecast, and an in-depth analysis of atmospheric phenomena, structure, composition, and dynamics. Cloud cover regulates the radiation balance in the Earth's climate system and is the most crucial factor of the solar energy flow to its surface [1–3]. Optical and microphysical models of the atmosphere are still imperfect. Their disadvantages include simplification due to low-level knowledge of atmospheric processes and phenomena. Account for the effect of high-level clouds (HLC) on the Earth's radiation budget in atmospheric climate models, is the most important unsolved problem [4, 5].

High-level clouds extending horizontally over thousands of kilometers, can cover up to a half of the Earth's surface [6, 7]. The HLC contribution to the greenhouse effect is significant despite their small optical thickness [8–10]. Parameters of the optical radiation transmission *via* HLC, are determined by their microstructure. It is characterized by the ice particle distribution in the cloud, shape, size, and spatial orientation, which, in turn, depend on meteorological conditions in the upper troposphere. In certain conditions, orientation of these particles becomes horizontal, thereby leading to an increased reflection coefficient and abnormal (specular) backscattering of optical radiation during zenithal lidar sensing. Atmospheric models, including the global atmospheric model of the European Center for Medium-Range Weather Forecasts, do not consider the HLC

https://doi.org/10.1007/s11182-024-03310-1

Published online: 06 January 2025

microstructure. Unlike droplet clouds, the particle size and shape of crystalline or mixed clouds need to be described accurately. In this regard, the effective radius is usually used based on the equality of one of the particle properties and some model spheres [1]. This simplification allows using Mie theory to calculate the HLC radiation parameters, although it is rather crude and negatively affects the accuracy of numerical weather and climate predictions.

Thus, experimental studies of such clouds are time-consuming and expensive. There are no contact methods of determination of the spatial orientation of particles in clouds, since they are violated during air sampling. A theoretical description of the interaction between the optical radiation and non-spherical ice particles is a complex, multi-faceted problem. Up-to-date tools and methods of processing experimental data, bring closer the creation of tools allowing to adequately link the atmospheric phenomena with meteorological conditions and predict the cloud microstructure. The polarization lidar can be used to compensate the inability to determine the particle orientation and other microstructure parameters (shape and size of ice particles) by a contact method.

Machine learning methods can be a powerful tool in solving climate problems, but they are not a universal solution for all problems and should be used in combination with other research and analytical methods to achieve the best results. In order for machine learning to solve a problem, it needs to be properly trained on a sufficiently large amount of high-quality data, and all factors influencing the problem must be considered. Machine learning in climate predictions may be ineffective if important climatic factors are excluded from the initial dataset, which may lead to incorrect conclusions.

The aim of this work is to study machine learning in atmospheric lidar sensing and processes important for addressing problems of atmospheric dynamics and physics. The developed software allows reconstructing the relation between meteorological parameters and HLC optical parameters and geometry.

Materials and methods

The high-altitude matrix polarization lidar (HAMPL) developed at the National Research Tomsk State University (Russia), is used to measure vertical profiles of atmospheric parameters. The HAMPL is able to perform all measurements necessary for the experimental determination of the vertical profile of all elements of the backscattering phase matrix (BSPM) of high-level clouds. The BSPM is a special case of the scattering phase matrix SPM for scattering angles close to 180 degrees [11]. Methods suggested in [12–15], allow to determine only individual BSPM elements, while others, if calculated, rely on the symmetry properties of these matrices.

In our approach, we use the systematically updated atmospheric lidar data storage from 2009 to 2024, comprising over 3000 series of measurements of atmospheric parameters [16]. The developed artificial neural network software is an application that employs machine learning algorithms to analyze weather observation data and predict optical and geometrical parameters of HLC. The proposed prototype can be used as a testing ground for various models to determine the most effective approach. It can also be used to demonstrate capabilities of the final product, such as evaluation of HLC boundaries and altitudes and BSPM elements based on the proposed software, as presented in [17–19].

The ERA5 reanalysis carried out by the European Centre for Medium-Range Weather Forecasts was utilized for the lidar data interpretation and analysis of meteorological conditions at the HLC altitude [20]. The ERA5 reanalysis provided regular and detailed vertical profiles of meteorological parameters, making it a valuable tool for complementing and refining aerological sounding data at a low frequency.

For the accuracy verification, the ERA5 dataset was compared to aerological sounding data collected for 5 years from 5 stations within a 500-km radius of Tomsk. It was shown that vertical profiles of meteorological parameters selected from the ERA5 data, corresponded to the actual conditions at the HLC altitude. That allowed using the ERA5 dataset for a detailed study of meteorological parameters in the upper atmosphere [19].

Table 1 Comparison of Operation on Full and Reduced Samples

Methods	BSPM elements	Sample reduction	Full dataset	Difference	Improvement, %
RF	m_{22}	0.032410	0.03398	–0.00157	–4.84
	m_{33}	0.073420	0.06021	0.01321	17.99
	m_{44}	0.118680	0.13480	–0.01612	–13.58
RF-PCA [15]	m_{22}	0.026780	0.02648	0.00030	1.12
	m_{33}	0.058430	0.05069	0.00774	13.25
	m_{44}	0.095730	0.09927	–0.00354	–3.70
RF-PCA [9]	m_{22}	0.027810	0.02820	–0.00039	–1.40
	m_{33}	0.064000	0.05373	0.01027	16.05
	m_{44}	0.097436	0.10653	–0.00909	–9.33

Results and discussion

In our recent research [17, 19], we used the subset of the data collected between 2016 and 2023 with focusing on the period when the HAMPL measurements were conducted in parallel. Simultaneously accumulated lidar signals to determine all the 16 BSPM elements corresponded to averaging over the same volume of the cloud or an ensemble of scattering particles. Measurements performed in parallel lidar signal accumulation, combined the data from more than 2500 series. Despite a significant number of lidar measurements, the number of measurement series from 2016 to 2020 amounted merely to 312, and recorded HLC were suitable for training neural networks. In terms of the experimental array expansion for the algorithm development and training, it was interesting to combine it with the data obtained between 2009 and 2016 in the form of sequential signal accumulation (over 1600 series). It was expected that the addition of new data would improve the algorithm performance, that would be a criterion for correct solutions. It was necessary to further seed the data array to improve the predictive capability of the model.

This study presents numerical calculations of BSPM elements with the enlarged data array in the sequential signal accumulation mode. Based on the ERA5 reanalysis, 124,512 vertical profiles of meteorological parameters with an hourly duty cycle for each parameter, are obtained for the period of 2009 to 2023. Each profile consists of more than 30 points on a non-uniform pressure grid. For standardization, all profiles are interpolated to a single altitude grid from 31 points. Incorporation of all these values in the neural network results in a substantial increase in the number of network parameters, which requires the data compaction to maximize information retention. The principal component analysis (PCA) is the classical method of reducing dimensionality. This method utilizes a singular value decomposition of the covariance matrix of the data. The spectrum of this decomposition characterizes components that carry the largest amount of information. According to [17], the HLC BSPM elements m_{22}, m_{33}, and m_{44} are subject to the analysis based on machine learning methods. In order to study the relationship between meteorological parameters and matrix elements, it is necessary to use the parameter value at the middle of the HLC altitude instead of the altitude profiles. The data array includes the following:

- determination of deferred sample for prediction (65 values in the data from February 15, 2020 to September 22, 2023 and 35 values from November 2, 2015 to January 22, 2016)
- determination of reduced sample for the first model training on the data from 2016 to 2023 obtained in lidar measurements with parallel signal accumulation
- determination of full sample for the second model training on the data from 2009 to 2016 obtained in lidar measurements with sequential signal accumulation

Machine learning algorithms are trained on reduced and full samples, followed by the result prediction on the deferred sample. The mean square deviation from the known coefficients for the deferred sample is

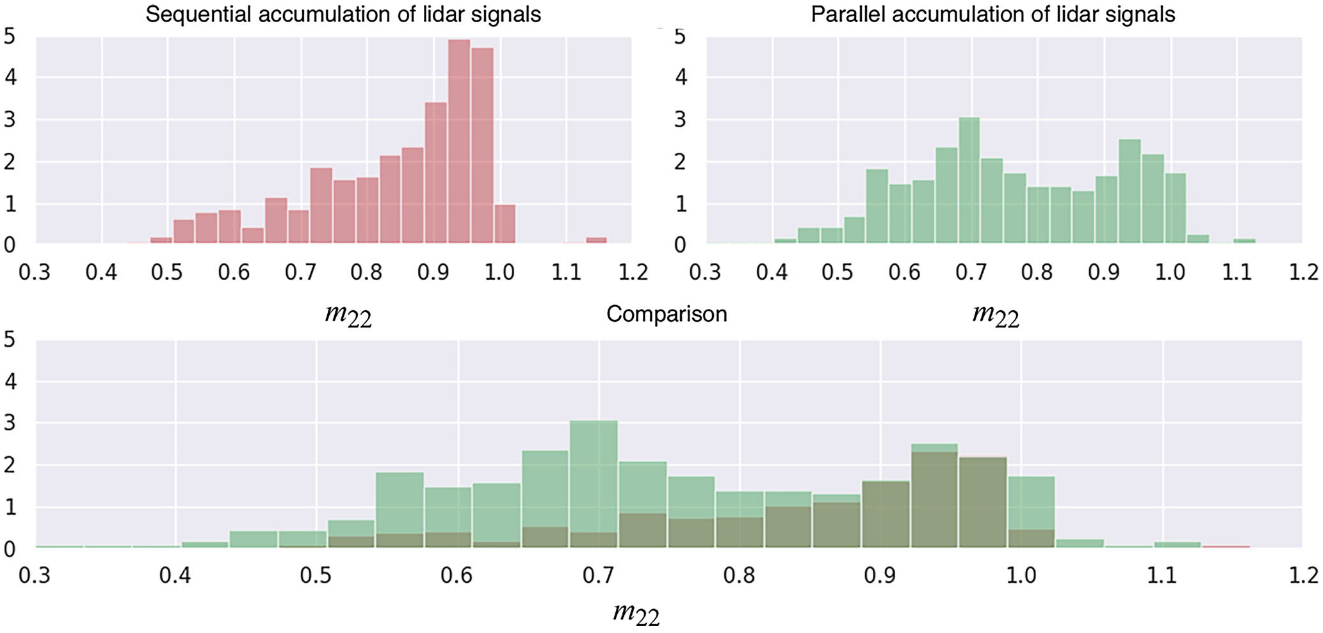

Fig. 1 Distribution diagrams of m_{22} element obtained by two different lidar measurement techniques

calculated as a metric for comparison. In machine learning, the random forest (RF) reduction method is used for dimensionality reduction using the principal component analysis (RF-PCA) with the number of components in parentheses. The results are summarized in Table 1. As can be seen from the table, the sample increase has a different effect on different BSPM elements. There are no significant differences for m_{22} element while for m_{33} element, the data are significantly improved. As for m_{44} element, the result is worse.

Next, it is necessary to determine whether the error increase for the validation sample is due to the independence of BSPM elements from weather conditions, suboptimal algorithm selection, and data representation, or if another error source relates to changes in the lidar measurement technique. For this, we check the distribution of BSPM elements in two samples, i.e., one is obtained from successive lidar measurements and the other—from parallel lidar measurements. The resulting distribution diagrams are shown in Figs. 1, 2 and 3. One can see significant differences between the two distributions of BSPM elements of HLC. In the case of sequential recording of lidar signals, a peak distribution is observed. This can be explained by the fact that weather conditions slightly change from measurement to measurement within this sample and, on average, all clouds have a common value for BSPM elements. The deviation from the average value depends on weather conditions. However, two peak values are observed for the parallel lidar signal accumulation method. Moreover, a comparison of the diagrams shows that one peak coincides with the peak obtained by the sequential method, while the second peak locates at a certain distance from it. For m_{33} element, this distance is the shortest.

Several peaks in the data obtained from the parallel lidar measurements, explain the deterioration of predictive capability of algorithms. This is because algorithms cannot adjust correctly both distributions, and in a fuller sample, behave as if they are trained only on a sequential sample. Accordingly, ignorance of the second peak data increases the prediction error. This also explains the improved prediction for m_{33} element compared to m_{22} and m_{44} elements, since the two peaks are next to each other, allowing algorithms to improve their performance.

However, the nature of the second peak remains unclear. There is a hypothesis that the second peak is associated with the detection of specular clouds, which requires further investigation, and the physical nature of the second peak will be discussed in a further paper.

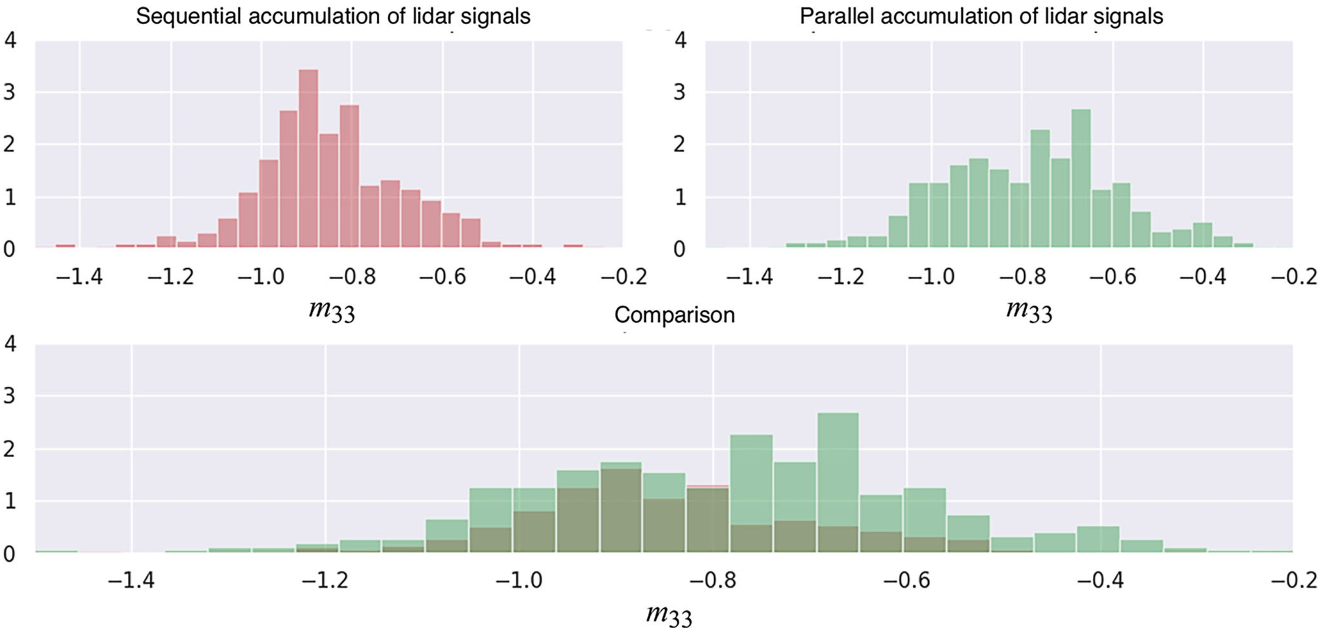

Fig. 2 Distribution diagrams of m_{33} element obtained by two different lidar measurement techniques

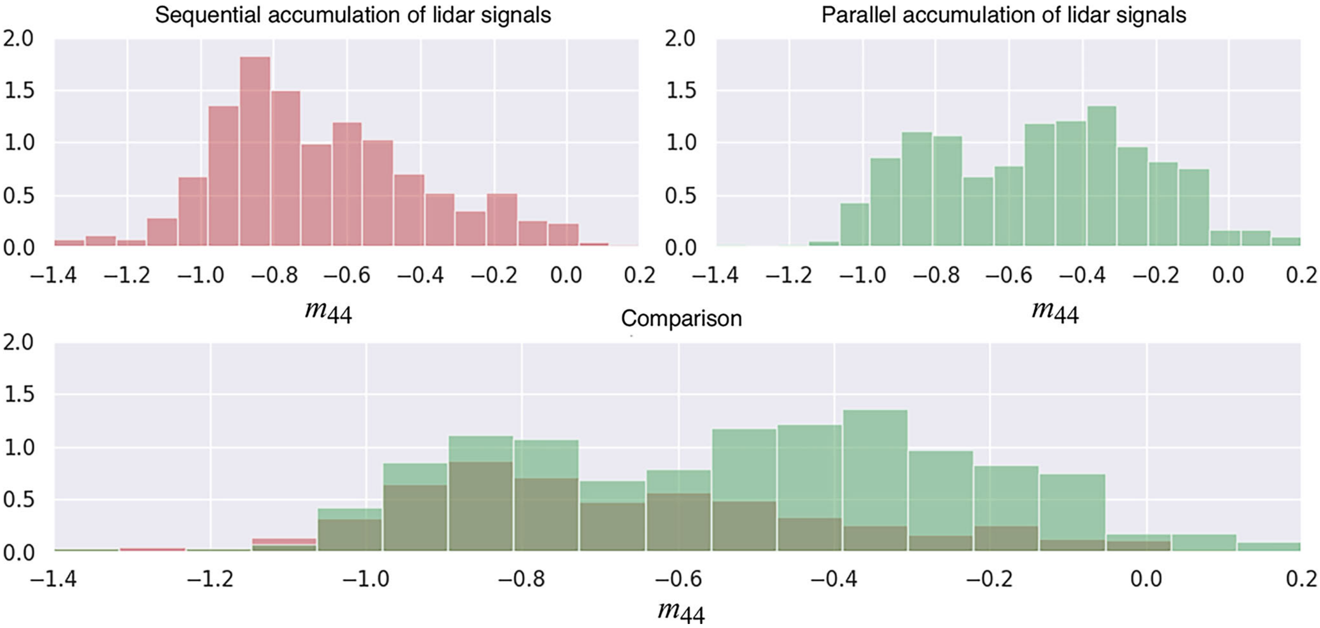

Fig. 3 Distribution diagrams of m_{44} element obtained by two different lidar measurement techniques

Conclusions

The proposed software utilized machine learning algorithms to analyze meteorological parameters and predict optical parameters and geometry of HLC. The prototype can be used to test various models and determine the most effective approach. It can also be used to demonstrate capabilities of the final product. The resulting tool allowed to preliminary evaluate BSPM elements and boundaries and altitudes of the HLC detection. A weak correlation was observed between HLC parameters and meteorological parameters. In order to determine this correlation, additional data are required to refine the results and consider experimental data, which we plan to collect soon.

Parallel lidar measurements demonstrated two maxima in the distribution of BSPM elements, which appeared to be of distinct origins. That assumption was based on the observation that the best predictions of BSPM elements were obtained by classification algorithms, rather than regression ones. The tendency of regression algorithms to produce constant values suggested that BSPM elements at different altitude likely had a common average value. However, there were heterogeneities, in which this value varied. It was assumed that these heterogeneities were specular clouds. Therefore, regression algorithms could not determine the dependence at constant values. However, the concentration of heterogeneities might depend on weather conditions. It was this dependence that classification algorithms could identify. Thus, a significant challenge emerged, namely to develop methods for determination of the heterogeneity concentration, understand their nature, and explore meteorological conditions under which they occur.

Funding This work was financially supported by Grant No. 075-15-2024-667 from the Ministry of Science and Higher Education of the Russian Federation. https://rscf.ru/project/23-79-00066.

Author Contribution Conceptualization I.S., O.K., I.B., and M.P.; preliminary analysis and lidar and meteorological data interpretation I.S. and I.B. O.K., and M.P.; software development by machine learning methods O.K., M.P., I.A., and D.R.; program-implemented comparison of BSPM element distribution I.A. and D.R.; data and measurement analysis I.B., M.P., O.K., and I.S.; writing—review and editing I.B., M.P., O.K., I.S., I.A., and D.R. All authors reviewed the final manuscript.

Conflict of interest The authors declare no conflict of interest.

References

1. Dmitrieva-Arrago, L.R., Trubina, M.A., Tolstyh, M.A.: Role of phase composition of clouds in forming high and low frequency radiation. Proc. Hydrometeorol. Res. Cent. Russ. Fed. **363**, 19–34 (2017)
2. Liou, K.N.: Influence of cirrus clouds on weather and climate processes: a global perspective. Mon. Weather. Rev. **114**, 1167–1199 (1986)
3. Wylie, D.P., Menzel, W.P., Woolf, H.M., Strabala, K.I.: Four years of global cirrus cloud statistics using HIRS. J Clim **7**, 1972–1986 (1994)
4. Reichardt, J., Reichardt, S., Lin, R.F., Hess, M., McGee, T.J., Starr, D.O.: Optical-microphysical cirrus model. J. Geophys. Res. Atmos. **113**, D22201:1–D22201:17 (2008)
5. Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (eds.): (2014)
6. Shanks, J.G., Lynch, D.K.: Specular Scattering in Cirrus Clouds. In: Lynch, D.K., Shettle, E.P. (eds.) DC, USA, Passive Infrared Remote Sensing of Clouds and the Atmosphere III, vol. 578, pp. 227–238. Bellingham (1995)
7. Heymsfield, A.J., Krämer, M., Luebke, A., Brown, P., Cziczo, D.J., Franklin, C., Lawson, P., Lohmann, U., McFarquhar, G., Ulanowski, Z., *et al.*: Cirrus clouds. Meteorol. Monogr. **58**, 2.1–2.26 (2017)
8. Mitchell, D.L., Finnegan, W.: Modification of cirrus clouds to reduce global warming. Environ. Res. Lett. 4, 045102:1–045102:8 (2009)
9. Storelvmo, T., Kristjansson, J.E., Muri, H., Pfeffer, M., Barahona, D., Nenes, A.: Cirrus cloud seeding has potential to cool climate. Geophys. Res. Lett. **40**, 178–182 (2013)
10. Vorobyova, V.V., Volodin, E.M.: Numerical simulation of influence on climate with the help of change of properties of high-level clouds in base of IVM RAS model. Proc. Hydrometeorol. Res. Cent. Russ. Fed. **363**, 5–18 (2017)
11. Samokhvalov, I.V., Kaul, B.V., Nasonov, S.V., Zhivotenyuk, I.V., Bryukhanov, I.D.: Backscattering light matrix of reflecting high-level clouds consisting of crystal mostly horizontally oriented particles. Atm. Ocean. Opt. **25**, 403–411 (2012)
12. 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. **45**, 4878–4887 (2006)
13. Hayman, M., Spuler, S., Morley, B., VanAndel, J.: Polarization lidar operation for measuring backscatter phase matrices of oriented scatterers. Opt. Express **20**, 29553–29567 (2012)
14. Volkov, S.N., Samokhvalov, I.V., Cheong, H.D., Kim, D.: Investigation of East Asian clouds with polarization light detection and ranging. Appl. Opt. **54**, 3095–3105 (2015)
15. Balin, Y.: S., Kaul, B.V., Kokhanenko, G.P.: Observations of specularly reflective particles and layers in crystal clouds. Atmospheric Ocean. Opt **24**(4), 293–299 (2011)

16. Ni, E.V., Kuchinskaya, O.I., Penzin, M.S., Bordulev, Yu.S., Bryukhanov, I.D., Doroshkevich, A.A., Samokhvalov, I.V.: A mechanism for predicting the optical and geometric characteristics of high-level clouds based on lidar and meteorological data. In: Proc. 29th Int. Sci. Conf. on Laser Information Technologies. Novorossiisk (2021), pp. 193–195
17. Kuchinskaya, 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. **14**(5), 1782 (2024)
18. Bryukhanov, I.D., Kuchinskaya, O.I., Ni, E.V., Penzin, M.S., Zhivotenyuk, I.V., Doroshkevich, A.A., Kirillov, N.S., Stykon, A.P., Bryukhanova, V.V., Samokhvalov, I.V.: Optical and geometrical characteristics of high-level clouds from the 2009–2023 data on laser polarization sensing in. Tomsk. Atmospheric Ocean. Opt. **37**(3), 343–351 (2024)
19. Kuchinskaya, O., Bryukhanov, I., Penzin, M., Ni, E., Doroshkevich, A., Kostyukhin, V., Samokhvalov, I., Pustovalov, K., Bordulev, I., Bryukhanova, V.: ERA5 reanalysis for the data interpretation on polarization laser sensing of high-level clouds. Remote Sens **15**, 109 (2023)
20. Copernicus climate data store. Available: https://cds.climate.copernicus.eu (accessed August 29, 2024)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

M. Penzin[1] · O. Kuchinskaia[1] · I. Akimov[1] · D. Romanov[1] · I. Bryukhanov[1] · I. Samokhvalov[1]

✉ M. Penzin
penzin.maksim@gmail.com

O. Kuchinskaia
olesia.kuchinskaia@cern.ch

I. Akimov
ima8908@mail.ru

D. Romanov
denrom@internet.ru

I. Bryukhanov
plyton@mail.tsu.ru

I. Samokhvalov
lidar@mail.tsu.ru

[1] National Research Tomsk State University, Tomsk, Russian Federation