Predictive capability of machine learning algorithms for reconstructing high-level cloud parameters based on lidar observations

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

Received: 4 September 2024 / Accepted: 6 October 2024

Abstract
The paper focuses on machine learning algorithms used to predict backscattering phase matrix (BSPM) elements of high-level clouds based on meteorological observations. Several machine learning methods, such as random forest, support vector, and linear regression, are used to detect the relationship between meteorological parameters and BSPM elements. It is shown that the random forest algorithm provides the most accurate predictions compared to other models. Despite a relatively small amount of the initial data, these methods have a good potential for their use in analyzing complex atmospheric interactions.

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

Introduction

In recent years, advances in machine learning have significantly enhanced the ability to predict complex physical phenomena through the analysis of large amounts of data [1]. This is particularly relevant in climatology, where weather patterns are influenced by numerous random factors [2]. Up-to-date information about weather changes is important for many areas of the human activity, e.g., protection from hazardous natural phenomena, anthropogenic environmental pollution [3, 4]. Long-term burning fossil fuels, deforesting, and interference in natural processes disrupt the climate system equilibrium. This leads to a disruption of the Earth's radiation balance, record growth in air temperature, abnormal precipitations, droughts, and other negative processes [5–7]. It is widely recognized that the cloud cover is one of the main sources of uncertainty in the prediction and understanding of climate changes. High-level clouds (HLC) have a large horizontal extent, and despite their low optical thickness (compared to stratiform clouds), they make a significant contribution to attenuation of direct solar radiation and enhancement of the greenhouse effect.

In climate models, radiative properties of high-level clouds are approximated, as the data on their microstructure are insufficient and primarily contain oriented crystalline ice particles [8, 9]. Diagnostics of the HLC microstructure is a challenging task, because in sampling air containing oriented aerosol particles, information about the particle orientation disappears. Therefore, only remote sensing methods based on optical radiation scattering by aerosol icy cloud particles, are suitable for detecting oriented particle ensembles in HLC.

The high-altitude matrix polarization lidar (HAMPL) developed in National Research Tomsk State University (Tomsk, Russia), provides remote sensing of the HLC microstructure parameters at a height of 15 km [10]. This requires lidar signal recordings for 4 polarization states of laser probing of polarization elements in 4 positions in the receiving channel [11, 12]. In general, to determine 16 BSPM elements, it is necessary to

 https://doi.org/10.1007/s11182-024-03311-0

Published online: 06 January 2025

solve a system of 16 integral equations. These inverse problems are quite difficult to solve. In the case of HLC containing non-spherical particles, the inverse problem becomes even more complicated, as there is no solution for the light scattering over particles with the shape differing from spherical. Due to these reasons, a traditional approach to the BSPM detection in crystalline clouds, based on solving the inverse problem of light scattering, faces sometimes inextricable difficulties. At the same time, neural networks are an effective tool to solve inverse problems and provide parallel training process, thereby facilitating relatively quick labor-intensive BSPM calculations based on the experimental data [13, 14].

The aim of this work is to study machine learning in atmospheric lidar sensing and processes important for addressing problems of ecology, atmospheric dynamics and physics. The most effective prediction algorithms are identified for BSPM elements, and their performance is evaluated in the context of various data compaction and feature extraction methods. The proposed methods of the parameter reconstruction of natural and artificial (contrail) high-level clouds under changing climate conditions and anthropogenic effects, utilize ground-based atmospheric sensing data and machine learning techniques. The software based on artificial neural networks, is created to predict BSPM elements of high-level clouds based on meteorological observations.

Machine learning in atmospheric optics is a significant step forward in the predictive ability and analysis of HLC optical parameters and geometry. This allows considering the complex interaction between meteorological parameters and reconstructing any dependence of atmospheric optical parameters. This is undoubtedly relevant for understanding the nature of their formation and evolution.

Materials and methods

The atmospheric lidar data storage from 2009 till now is used to evaluate the predictive capability of parameter recovering algorithms of high-level clouds based on the lidar data and machine learning. The lidar data storage comprises over 3000 series of measurements of atmospheric parameters and is systematically updated. Lidar measurements are the most promising in addressing issues of operational monitoring and management of atmospheric conditions, as they determine vertical profiles of optical, microphysical, and meteorological parameters of the scattering medium in real time.

Lidar observations are noninvasive testing which has no adverse effect on the environment. The analysis of laser radiation parameters following its interaction with cloud particles, provides the evaluation of scattering volume properties, including microphysical parameters such as size, shape, and particle orientation. With the high spatial (37.5–150 m) and temporal (0.1 s) resolution, the lidar identifies localized inhomogeneities characterized by the gradient of optical and microphysical characteristics [12]. Systematization and analysis of lidar measurements provide a deeper understanding of the HLC formation and evolution.

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 [15]. 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 [16].

Results and discussion

In our recent research [17, 18], we used machine learning methods to determine the empirical relationship between meteorological and HLC parameters. The relation between meteorological parameters and the HLC

altitude was investigated to determine the HLC boundaries and predict BSPM elements based on meteoritical conditions. It was demonstrated that only m_{22}, m_{33}, and m_{44} elements of the BSPM main diagonal depended on weather conditions. Specialized software was employed for the data processing and analysis with a high accuracy and reliability.

In this study, we determined the most effective algorithm for predicting BSPM elements of clouds. For this, several conventional machine learning methods were evaluated, including random forest, linear regression, support vector regression, and principal component analysis (PCA). The data transformation was performed using normalization techniques [19]. These models were selected due to their simple configuration, resistance to overfitting (due to regularization), and dealing with limited data volumes.

To determine the best model hyperparameters, the data were split into training, test, and validation sets. These hyperparameters resulted in the lowest mean squared error (MSE) in the test set, calculated from Eq. 1. The training set consisted of 247 observations gathered between March 22, 2016, and April 9, 2019. The test set included 49 observations gathered between April 16, 2019, and March 17, 2020 as well as from September 22 to November 30, 2023. The validation set was used to assess the final model quality, rather than for the model training or hyperparameter selection. It consisted of 49 additional observations gathered between March 18 and July 19, 2023.

$$\mathrm{MSE} = \frac{1}{N} \cdot \Sigma \left(y_{pred} - y_{dat}\right)^2 \tag{1}$$

Standard hyperparameter selection techniques were employed to minimize the MSE between the predicted and measured values of the test dataset. Specifically, for the random forest model, the primary hyperparameter included a number of trees in the ensemble. The best parameter was found to be 300. Regularization was not applied to the linear regression model and thus no hyperparameter was adjusted for it. For the support vector regression (SVR), the core hyperparameter was the kernel. Therefore, different kernels were tested, and the radial bas is function kernel was found to produce the best results in the test set:

$$K\left(x, x'\right) = \exp^{\gamma|x-x'|2}, \tag{2}$$

where $|x–x'|^2$ is the squared distance between vectors x and x' calculated by Euclidean metrics, γ is the parameter regulating the kernel width.

The data preprocessing, including the PCA and normalization of the feature space, is conducted to improve the model performance. The number of principal components varies between 10 and 30, and 20 performing bests. The average value of these components is subtracted from each feature, and resulting values are then subjected to the standard deviation. This results in each feature, having zero average value and standard deviation of one.

The best algorithm for predicting BSPM elements based on atmospheric parameters, is presented in Table 1. We select the dataset of optimized 345 observations from 2016 to 2024. 296 of them are used as the training set and 49—as the validation set. After the model training, validation set predictions are made, and the MSE quality metric is calculated.

The best results for m_{33} and m_{44} elements are archived using the random forest in conjunction with the PCA. For m_{22} element, the SVR with the feature field transformation techniques (PCA and standardization) demonstrates the highest accuracy.

Figure 1 contains scattering diagrams corresponding to the best outcomes, with the actual predicted and measured values plotted on the horizontal and vertical axes, respectively. A straight line is expected at an ideal definition of BSPM elements, as the predictive value coincides with the measured one. However, due to the presence of noise or weak dependence on the input data, a difference between the predicted and measured values is observed.

https://doi.org/10.1007/s11182-024-03311-0

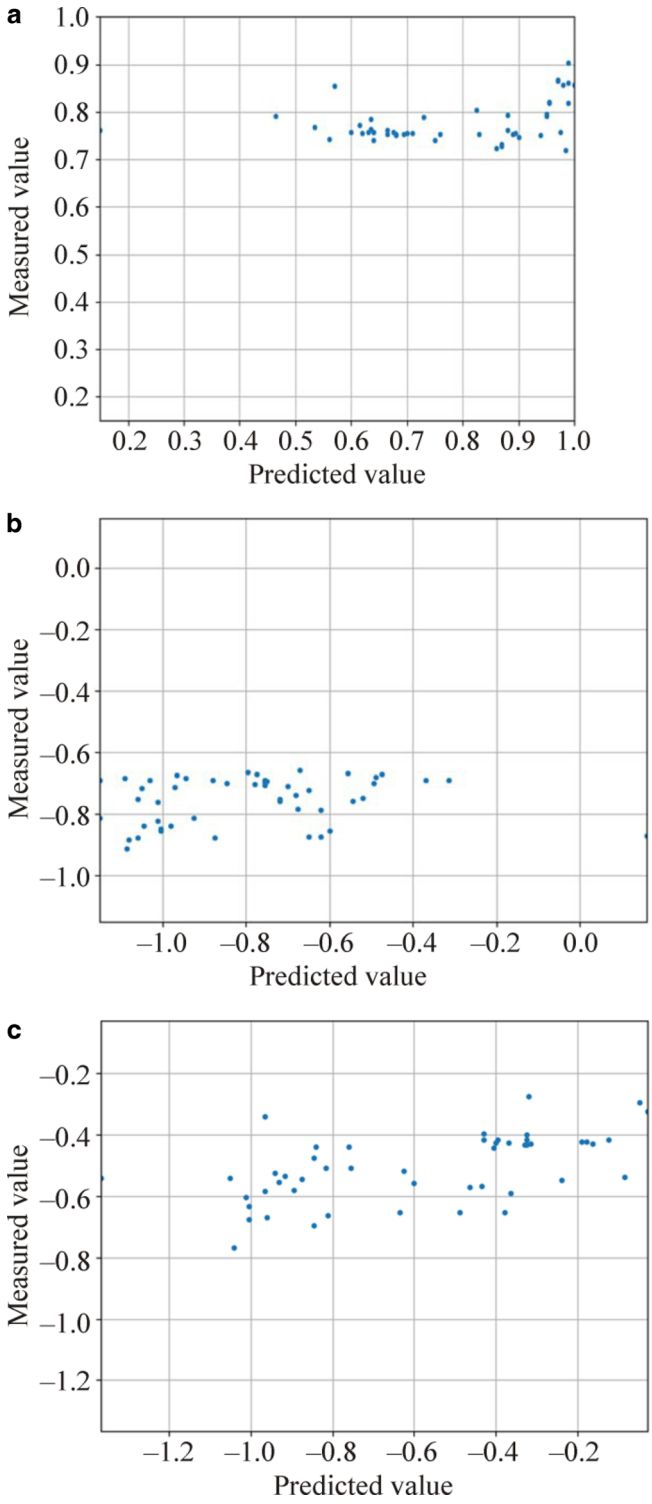

Fig. 1 Scattering diagram of BSPM elements: **a** m_{22} element predicted by SVR, **b** m_{33} element predicted by RF+PCA(20)+scale, **c** m_{44} element predicted by RF+PCA(20)+scale

In all cases, we observe a certain degree of variation, which indicates that it is feasible to estimate BSPM elements using machine learning tools and meteorological parameters. The error is however significant enough to limit the applicability of these results. This issue can be addressed through additional specific information about the dynamics of profile changes over time and anthropogenic influences. The range extension of the experimental data will provide more conclusive results.

Table 1 Algorithm Comparison

Methods	HLC BSPM elements	MSE
RF	m_{22}	0.03389
	m_{33}	0.08311
	m_{44}	0.13201
RF + PCA(20)	m_{22}	0.03023
	m_{33}	0.06572
	m_{44}	0.08723
LR + PCA(20) + scale	m_{22}	0.02777
	m_{33}	0.10220
	m_{44}	0.13197
LR + PCA(20)	m_{22}	0.02891
	m_{33}	0.06998
	m_{44}	0.11223
SVR	m_{22}	0.02758
	m_{33}	0.06929
	m_{44}	0.09901
SVR + PCA(20) + scale	m_{22}	0.04750
	m_{33}	0.13139
	m_{44}	0.14655

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. 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. The obtained results were analyzed to determine the most suitable algorithm for predicting BSPM elements. It was found that in conjunction with the PCA, the random forest yielded the highest accuracy for m_{33} and m_{44} elements. The support vector regression with the PCA-transformed feature field and standardization demonstrated the best accuracy for m_{22} element.

Despite the model optimization, the overall trend of predictions remains low-variable and requires continued improvement. The prediction quality can be improved by the transformation of target variables, such as Box-Cox or Yeo-Johnson transforms. Additionally, Kolmogorov-Arnold networks may also be explored [20]. It is thus necessary to employ different models for the accurate prediction of BSPM elements, which must be optimized for each aspect. This approach will improve the prediction accuracy and reliability of various BSPM elements based on meteorological parameters.

Funding Partial financial support was received from the Russian Science Foundation (Grant No. 24-72-10127) for studying the relationship between optical, microphysical, and geometric parameters of high-level clouds with horizontally oriented icy cloud particles and meteorological conditions leading to their formation and evolution. And partial financial support was received from the Russian Federation government (Agreement No. 075-15-2024-667) for the software development in the field of the BSPM element distribution using sequential and parallel lidar signal accumulation using the high-performance computing cluster of National Research Tomsk State University.

Conflict of interest The authors declare no conflict of interest.

References

1. Bochenek, B., Ustrnul, Z.: Machine learning in weather prediction and climate analyses—applications and perspectives. Atmosphere **13**, 180 (2022). https://doi.org/10.3390/atmos13020180
2. Wilks, D.S.: ISSN. Stat. Methods Atmospheric Sci. Int. Geophys. **100**(627), 74–6142 (2006)
3. Donnelly, J., Abolfathi, S., Pearson, J., Chatrabgoun, O., Daneshkhah, A.: Gaussian process emulation of spatiotemporal outputs of a 2D inland flood model. Water Res. **225**, 119100 (2022)
4. Kim, I., Kim, B., Sidorov, D.: Machine learning for energy systems optimization. Energies **15**, 4116 (2022)
5. 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 **8451812**(2022), (2022)
6. 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. **220**, 103752 (2021)
7. Nourani, V., Khodkar, K., Paknezhad, N.J., Laux, P.: Deep learning-based uncertainty quantification of groundwater level predictions. Stoch. Environ. Res. Risk Assess. **36**, 3081–3107 (2022)
8. 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.): (2013)
9. 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)
10. Scientific and Technological Infrastructure of the Russian Federation. Radiophysical Complex: High-Altitude Polarization Lidar for Atmospheric Sensing and Tomsk Ionospheric Station "LIDAR-IONOSONDE". Available: https://ckp-rf.ru/catalog/usu/73573 (accessed August 30, 2024)
11. Kaul, B.V., Volkov, S.N., Samokhvalov, I.V.: Studies of ice crystal clouds through lidar measurements of backscattering matrices. Atmospheric Ocean. Opt. **16**(4), 325–332 (2003)
12. Kaul, B.V., Samokhvalov, I.V., Volkov, S.N.: Investigating particle orientation in cirrus clouds by measuring backscattering phase matrices with lidar. J. Appl. Opt. **43**(36), 6620–6628 (2004)
13. Lee Wilson, Tiong T. Goh, William Yu Chung Wang: Big data in climate change research: opportunities and challenges. IJEA **4**(2), 1–14 (2020)
14. Cao, L.: Data Science: A Comprehensive Overview. ACM Computing Surveys (CSUR) **50**(3), 1–42 (2017)
15. Copernicus climate data store. Available: https://cds.climate.copernicus.eu (accessed August 29, 2024)
16. Kuchinskaia, 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)
17. 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. **14**(5), 1782 (2024). DOI:0.3390/app14051782
18. Bryukhanov, I.D., Kuchinskaia, 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. Aurélien Géron: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Brown, K. (ed.) O'Reilly Media, Inc., Sebastopol, CA (2019)
20. Ziming, Liu, Wang, Yi., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Max Tegmark Kan: Kolmogorov-Arnold Networks. arXiv:2404.19756 (2024). https://doi.org/10.48550/arXiv.2404.19756

Authors and Affiliations

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

✉ D. Romanov
denrom@internet.ru

I. Akimov
ima8908@mail.ru

M. Penzin
penzin.maksim@gmail.com

O. Kuchinskaia
olesia.kuchinskaia@cern.ch

I. Samokhvalov
plyton@mail.tsu.ru

I. Bryukhanov
lidar@mail.tsu.ru

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