Development of machine learning models for predicting average annual temperatures

Kirill Mukhin[1], *Viktoriya* Erofeeva[1,2], and *Zhanna* Zhukova[2]

[1]Peoples' Friendship university of Russia named after Patrice Lumumba, 115093 Moscow, Russia
[2]Moscow Technical University of Communications and Informatics, 111024 Moscow, Russia

Abstract. This study assesses machine learning models for predicting Antarctica's average annual temperatures, addressing the challenge of accuracy in remote and variable climatic conditions. Four models were compared: linear regression, random forest regressor, decision tree regressor, and gradient boosting, utilizing data from diverse Antarctic stations. Results indicate the superiority of specific models tailored to individual stations, with the random forest model demonstrating exceptional performance across most metrics. This emphasizes the significance of geographical specificity in improving climate prediction accuracy. The research underscores machine learning's potential in climate change forecasting, advocating for tailored approaches in environmental modeling.

1 Introduction

The study of climate change, particularly within regions with extreme weather conditions such as Antarctica, stands as an important aspect of modern Earth science. Antarctica, occupying a key position in the global climate system, affects ocean currents [1] and atmospheric processes at the global level. Moreover, the prediction of air temperature in Antarctica possesses an important role in environmental, glaciological and climatological processes. Predicting temperatures in this region is not only a scientific interest, but also a necessity for understanding future climate scenarios on the planet. The development of machine learning technologies introduces new prospects for the accuracy and reliability of forecasts of climatic parameters. This area of machine learning application has gained the importance due to the difficulties in achieving high accuracy of temperature prediction. Specifically, it has been proved that the instability of temperature datasets adheres to intricate, long-range correlation, demonstrating nonlinear behavior [2].

Besides, there are several other problems with predicting temperatures in Antarctica. First, network of weather stations is not stable [3] due to the geographical remoteness of Antarctica, and it is not possible to predict the temperature across the whole territory. Second, the use of machine learning methods is limited by low time resolution, for example, annual or monthly temperature averages [4].

This study presents the results of the development and comparison of four machine learning models for predicting temperatures in Antarctica: regression models, random forest regressor, decision tree regressor and gradient boosting. Emphasis is placed upon comparing the model performance adapted for specific stations and general models designed for use in

a wide range of locations. Encompassing three categories of stations—proximal to the pole, coastal, and intermediary zones between the pole and the coast—the study facilitates to assess the impact of geographical location on the accuracy of predictions.

The primary objective of this study is not only to demonstrate the prospects of machine learning in predicting temperatures in the extreme conditions of Antarctica, but also to identify the most effective modeling approaches, considering the diversity of climatic conditions in different parts of the continent. By analyzing specialized and general models, this work aims to contribute to the optimization of machine learning strategies for climate research.

2 Methodology

In order to develop machine learning models for predicting changes in average annual temperatures in Antarctica, data from available sources is collected. Our dataset encompasses observations from 3 different types of stations. There is 1 station close to the pole – Amundsen-Scott station, 4 stations on the coast on different sides of the mainland – Mirny station, Baird station, Vernadsky station and McMurdo station, as well as 1 more station inland – Vostok station. Such a comprehensive array of stations allows us to evaluate the possibilities of temperature forecasting in different conditions.

The research and modeling plan comprises several key stages:

1. Data collection from open sources.
2. Data preparation and feature engineering.
3. Machine learning model construction: development of predictive models.
4. Model evaluation: assessment of model accuracy.

2.1 Feature engineering

During the feature engineering stage, we decided to extract lag features from the average annual temperature dataset. These lag features serve to inform both the models and analysts regarding the directional tendencies of the function x(t)—whether it exhibits growth or decline—and a moving average to delineate the overarching temporal trend. These features are selected individually for each model, allowing us to improve their quality.

2.2 Application of machine learning algorithms

This study employed four primary machine learning methodologies to forecast temperatures across different Antarctic locations: linear regression, decision tree regressor, random forest regressor, and gradient boosting, implemented via the CatBoost library. The selection of these methods stemmed from their widespread adoption and efficacy in forecasting tasks. A careful selection of hyperparameters was carried out for each model to optimize its performance.

Linear regression.

Linear regression is a basic machine learning method that assumes a linear relationship between the input variables and the target variable, used here to establish a simple relationship between the characteristics of climate data and temperature. Hyperparameter selection focused primarily on regularization method to prevent overfitting.

Decision tree regressor.

The decision tree models predictions uses a tree structure, where each node represents a decision point based on a specific attribute, and branches represent the outcomes of these decisions. This method is useful for interpreting and understanding what factors influence

the predicted temperature. Hyperparameters such as the depth of the tree and the minimum number of samples to split the node have been configured to achieve the best performance of the model.

Random forest regressor.

The random forest regressor is an ensemble method based on aggregating the results of multiple decision trees to improve the accuracy and stability of forecasts. This approach allows you to effectively manage retraining and improve the accuracy of forecasting. In the process of selecting hyperparameters, parameters such as the number of trees in the forest and the depth of the trees were considered.

Gradient boosting (CatBoost).

Gradient boosting is a machine learning technique that builds a prediction model in the form of an ensemble of weak predictive models, usually decision trees. CatBoost is one of the advanced implementations of gradient boosting, characterized by the efficiency of processing categorical data and reducing retraining. The selection of hyperparameters for CatBoost included setting the learning rate and the depth of the trees.

Each of these machine learning methods demonstrates its unique advantages in predicting temperatures in Antarctica. Careful selection of hyperparameters made it possible to adapt each model to the characteristics of the data as much as possible, ensuring high accuracy and reliability of forecasts. Next, the results of using different models will be compared to determine the most effective approach in different climatic conditions of Antarctica.

2.3 Model accuracy evaluation

Model accuracy was assessed by employing the Mean Absolute Error (MAE) metric—a straightforward indicator measuring the average absolute discrepancy between actual and predicted values, thereby offering intuitive insight into forecast quality.

2 Results

At the first stage, individual models were constructed for each station, each tailored with specific parameters (Fig. 1). Based on the results obtained, it is possible to make conclusions about a positive result, since the average absolute error at all stations does not exceed 1 degree. Notably, the random forest regressor model consistently outperformed other models. In 4 out of 6 stations, it showed the best result and never turned out to be the worst model, so we can conclude that it is best to use a random forest regressor model for temperature prediction algorithms at each station individually. The same figure shows that the decision tree regressor model showed the best result in the general model. At the same time, in only 1 case it showed a better result than the individual model, and based on the results in Figure 2, it can be concluded that for a more accurate result it is better to use individual models for each station.

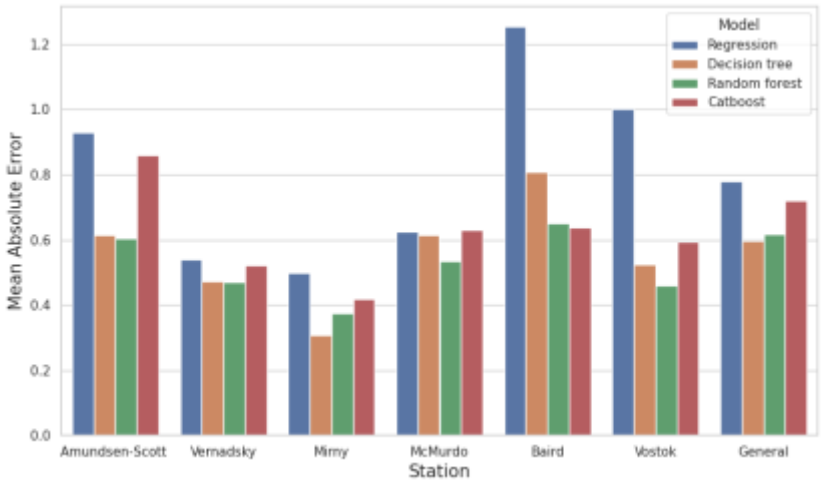

Fig. 1. Comparaison of the results of individual models.

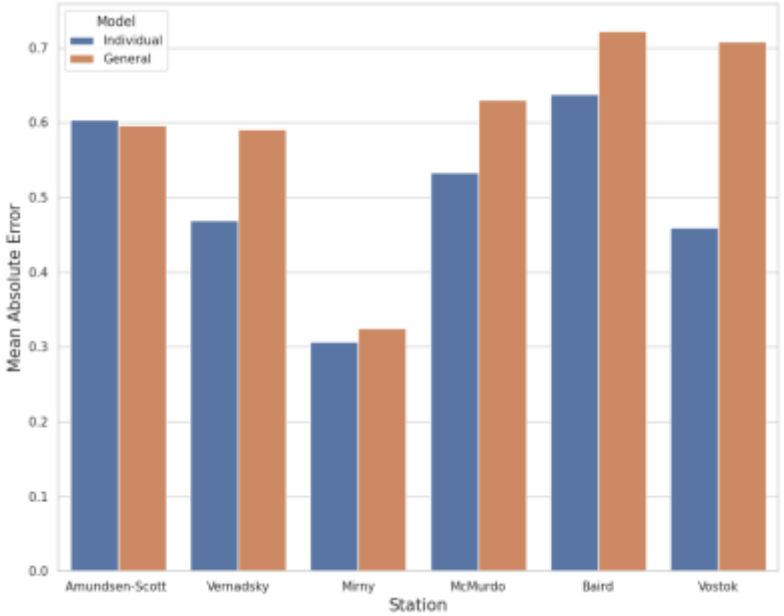

Fig. 2. Comparison of the results of the best individual models with the best general model

3 Conclusions

In conclusion, we developed and compared four distinct machine learning methodologies. Based on a comparison of the performance of models adapted for specific stations and general models designed for use in a wide range of locations, it became evident that individual models demonstrate better results in temperature prediction accuracy. This emphasis on individual models underscores the significance of considering the unique climatic characteristics of each specific location in Antarctica, thereby facilitating more reliable and precise forecasts. The results of this study confirm the potential of machine learning as a powerful tool for predicting climate change in Antarctica. Futhemore, the findings underscore the necessity for

an individualized modeling approach, which can help improve the accuracy and reliability of climate change forecasts different parts of the world.

Looking ahead, researchers will need to explore the adaptability of various machine learning approaches to other aspects of research [5-6], as well as develop new techniques to improve model performance. Additionally, an important area of work is to improve the methods of data collection and processing, which can further contribute to improving the effectiveness of forecasting.

References

1. Q, Li, J. Clim. **36(11)**, 3571–3590 (2023).
I. Bartos, I. Jánosi, Nonlinear Process. Geophys. **13**, 571–576 (2006)
2. M. A. Lazzara, G. A. Weidner, L. M. Keller, J. E. Thom, J. J. Cassano, Bull. Amer. Meteor. Soc. **93**, 1519–1537 (2012)
3. Y. Wang, S. Hou, Prog. Nat. Sci. **19**, 1843–1849 (2009).
4. Y. Mayorova, V. Glebov, V. Erofeeva, S. Yablochnikov, B. Laver, E3S Web Conf. **169**, 5 (2020)
5. V. V. Erofeeva, V. V. Vasenev, Springer Geogr. **2020**, 52–57 (2020)