ISSN 0027-1349, Moscow University Physics Bulletin, 2023, Vol. 78, No. 7, pp. 210–216. © Allerton Press, Inc., 2023.
Russian Text © The Author(s), 2023, published in Vestnik Moskovskogo Universiteta, Seriya 3: Fizika, Astronomiya, 2023, Vol. 78, No. 7, 2350105.

MACHINE LEARNING IN NATURAL SCIENCES

SMAP Sea Surface Salinity Improvement in the Arctic Region Using Machine Learning Approaches

A. S. Savin[1,2*], M. A. Krinitskiy[1,2], and A. A. Osadchiev[1,2]**

[1] *Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia*
[2] *Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia*

Received August 21, 2023; revised October 5, 2023; accepted October 5, 2023

Abstract—Sea surface salinity (SSS) is a key physicochemical characteristic of the ocean that plays a significant role in describing the climate. Routine SSS retrieval algorithms exploiting remote sensing data have been developed and validated with high precision for typical regions of the World Ocean. Their effectiveness is worse in the Arctic though. To address this limitation, in this study, we employ machine learning (ML) techniques to enhance the quality of standard algorithms. We evaluate a few ML models, ranging from classical methods that process vector features, provided by standard Soil Moisture Active Passive (SMAP) satellite salinity algorithms, to deep artificial neural networks that combine vector features with two-dimensional fields extracted from the ERA5 reanalysis. We validate these models using *in situ* the data collected by the Shirshov Institute of Oceanology RAS during the expeditions to the Barents, Kara, Laptev, and East Siberian seas from 2015 to 2021. The results of the study indicate that the SMAP sea surface salinity standard product is improved in these regions. The ML models developed in this study make it possible to further study the Arctic region using enhanced sea surface salinity maps.

Keywords: sea surface salinity, machine learning, deep learning, artificial neural networks, multilayer perceptron, convolutional neural networks, CatBoost, Random Forests, Gradient Boosting, SMAP, Arctic region

DOI: 10.3103/S0027134923070299

1. INTRODUCTION

Sea salinity is one of the crucial characteristics of physical processes occurring in the World Ocean and is currently considered to be one of major climate variables [1]. Combined with temperature, salinity determines the global currents system that control the circulation of the entire World Ocean [2]. Sea salinity is influenced by numerous processes, and therefore serves as an indicator characterizing the physical and chemical processes occurring on Earth [3]. For its part, the salinity of ocean water affects, for example, its density and other characteristics, so measuring salinity allows obtaining information not only about the water itself and its internal structure, but also about the interaction between ocean and atmosphere.

Nowadays, satellite measurements are actively developed and used to obtain sea surface salinity (SSS) value. The missions Soil Moisture and Ocean Salinity (SMOS, since 2009), Aquarius (since 2011), and Soil Moisture Active Passive (SMAP, since 2015) provide information about the salinity field through satellite measurements. Microwave radiometers operating in the L-band (1.4 GHz) installed on these satellites allow retrieving SSS values based on microwave radiation data [4].

Standard numerical algorithms used for SSS retrieval values from microwave radiation data have been developed and verified with high accuracy for the most typical ranges of temperature and salinity in the World Ocean [4]. However, these standard algorithms have significantly lower accuracy when retrieving salinity in the Arctic Ocean, which is characterized by low temperatures and is influenced by substantial freshwater runoff [5–9].

The aim of this study is to improve the standard SMAP SSS algorithm using machine learning approaches for the Russian Arctic seas. The algorithm is trained and verified using *in situ* salinity measurements obtained from numerous expeditions conducted from 2015 to 2021 in the Barents, Kara, Laptev, and East Siberian Seas.

*E-mail: savin.as@phystech.edu
**E-mail: krinitsky@sail.msk.ru

2. DATA

2.1. SMAP Data

NASA/RSS SMAP Salinity satellite data, version 4.0, level 2C, distributed by the National Aeronautics and Space Administration (NASA) Physical Oceanography Distributed Active Archive Center (PO.DAAC), were used to estimate SSS. SMAP brightness temperature in vertical and horizontal polarizations measured by the satellite and SSS SMAP—the standard SMAP product for science applications—are used, as well as characteristics provided by other sources synchronized with satellite measurements. These include sea surface temperature from the Canadian Meteorological Center, sea surface wind speed and its direction from Cross-Calibrated Multi-Platform wind vector analysis and mean solar flux from NOAA National Centers for Environmental Prediction. Besides, the zenith and azimuth Sun angles for the observation point, the land fraction weighted by the antenna gain pattern, land fraction within the footprint, and the sea ice fraction weighted by the antenna gain pattern are used due to region specificity. All the satellite-measured and ancillary data are aligned on a unified spatial-temporal grid and distributed jointly by NASA PO.DAAC.

2.2. ERA-5 Data

In addition to the data distributed by NASA PO.DAAC, this study also utilizes certain ERA-5 climate data on the near-surface atmospheric conditions. These are data from the European Centre for Medium-Range Weather Forecasts, distributed by the Copernicus Climate Change Service. These include mean sea-level pressure, air temperature at a height of 2 meters above the surface, as well as zonal and meridional wind components at a height of 10 meters. The consideration of these parameters is explained by their significant influence on the satellite-observed signal, which can lead to potential distortions of the standard algorithm results.

2.3. In Situ Data

The *in situ* data used in this study were collected during the expeditions conducted by the P.P. Shirshov Institute of Oceanology, Russian Academy of Sciences, from July to October in 2015-2021. In total, more than 1,3 million measurements of SSS were collected. The coverage of measurements in the Arctic Ocean is shown on the map in Fig. 1.

Distribution of the number of SSS data is shown in Fig. 2. The majority of the collected data refer to the open ocean, where SSS values range from 24 psu and above. This range will be referred to as "high" salinity values. "Low" salinities, on the other hand, correspond mainly to areas where a strong influence of river runoff is observed, with SSS values below 15 psu. Salinities ranging from 15 to 24 psu will be referred to as "medium" class, corresponding to the mixing zone of the waters from the two types described above.

3. METHODOLOGY

3.1. Data

In this study, the improvement of the standard SMAP product is achieved using machine learning (ML) approaches. The ML models use a feature set consisting of 13 variables, with 12 of them being features for the standard SMAP algorithm, and one being the SMAP default product itself. These variables include, primarily, the brightness temperature values measured by the satellite in vertical and horizontal polarizations. Additionally, other features measured by ancillary sources, as mentioned earlier, are also included. The selection of these parameters is motivated by the following reasons.

Firstly, the complexity of the problem being addressed is a primary factor. The improvement of SSS using SMAP data is carried out for the entire range of salinity values from 0 to 35 psu, and relying solely on brightness temperature data proves to be insufficient. Secondly, the characteristics of the considered region significantly complicate the task. Ice drift, formation, and melting, combined with high land presence, considerably reduce the accuracy of standard algorithms [cite articles]. To take it into account, features that characterize the proportion of land and ice in the satellite image are considered. Additionally, the quality of the satellite image is affected, in part, by the Sun angles at the considered location. The features that describe Sun angles are intended to serve as a "quality flag" for the ML models. In some kinds of ML models two-dimensional ERA-5 data, including sea level pressure, temperature, and wind speed and direction, was also used to estimate SSS.

In situ data in this study are used to validate the ML models. The *in situ* data are matched to the satellite data based on spatial and temporal proximity. The distance between the matched satellite and *in situ* measurement points does not exceed 10 km, and the time difference is not more than 3 hours [10]. This is due to the variability of the situation in the considered region and the feature representation used. The data are split into training and testing datasets for model validation based on dates of the year.

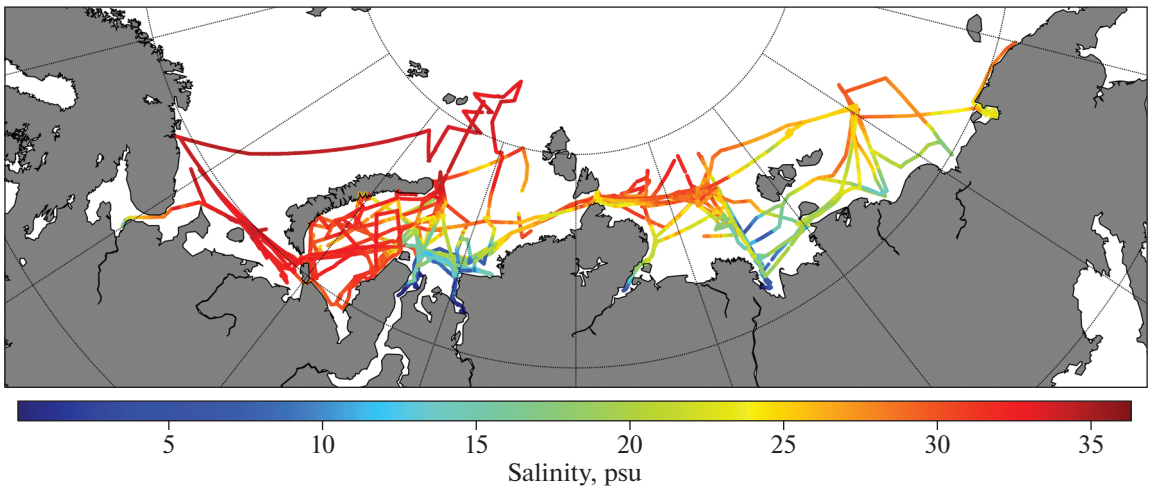

Fig. 1. *In situ* data spatial distribution. Measurements conducted in marine missions of Shirshov Institute of Oceanology of Russian Academy of Sciences in 2015–2021.

3.2. ML Models

Machine learning, which has been previously applied to different remote sensing tasks [11, 12], including similar ones [10], is the main method used in this research. Both classical models and artificial neural networks are considered. The first examined classical ML model is the Random Forest model [13], which is often used in regression and/or classification tasks. In this study, the Random Forest model implemented in the scikit-learn [14] framework is used. The second classical ML model considered is Gradient Boosting. The CatBoost Gradient Boosting model, developed by Yandex, is used in this research [15]. The search for optimal hyperparameters for the Random Forest and Gradient Boosting models was performed using the optuna framework [16].

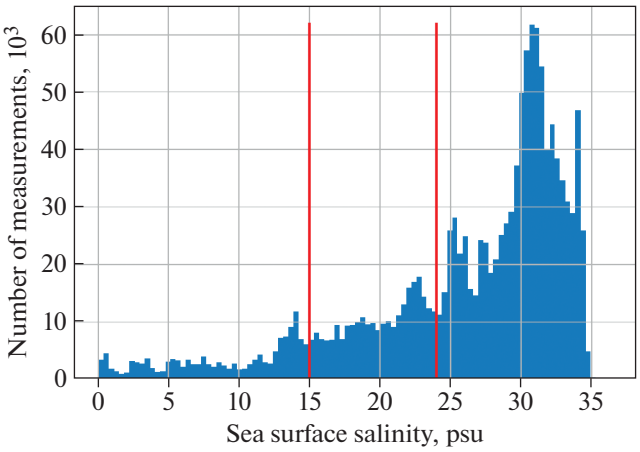

Fig. 2. Number of *in situ* measurements. We also present here the three classes of salinity values, i.e., "low", "medium" and "high" salinity, visually divided here with red vertical lines.

All the mentioned classical models are considered in two configurations: "single-level" and "two-level". The "single-level" models are used to solve the regression task predicting the SSS for the entire spectrum of the provided *in situ* measurements. On the other hand, the "two-level" models have a more complex configuration: first, a classification task is solved to classify the salinity into "low," "medium," and "high" classes, as described earlier, and then, based on the predicted class, a regression task is performed to predict SSS.

In addition to the classical models described above, various types of artificial neural networks (ANN) are considered in this research. The first model is a fully connected neural network, also known as a Multilayer Perceptron (MLP), which is the most typical ANN configuration. In the implementation of this model, a vector feature representation is used, which consists of the 13 satellite variables. Similar to the classical models, a "two-level" analog is also considered in the neural network approach. However, unlike the aforementioned solutions, where classification and regression tasks are, in fact, solved sequentially, in the proposed algorithm, the class values obtained from the satellite variables themselves are used as input features for estimating SSS.

The last examined configuration of artificial neural networks allows for the consideration of not only satellite measurements but also two-dimensional fields of pressure, temperature, and wind in the area of the observation point. Fig. 3 demonstrates the architecture of such a model.

The features of the two-dimensional description are derived using Convolutional Neural Networks (CNN). To achieve effective feature extraction, the convolutional part of the overall neural network is pre-trained using an autoencoder approach on the

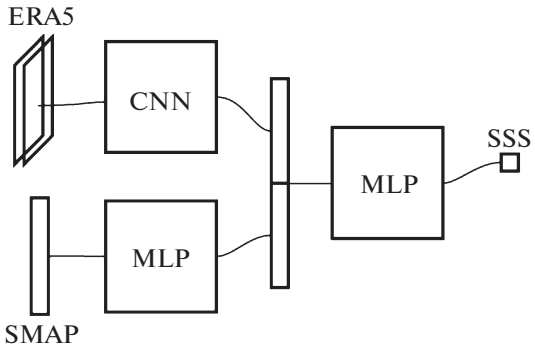

Fig. 3. Composite neural model considering SMAP and ERA-5 features.

data from July to October 2000-2019. The data for pre-training are randomly selected uniformly across time and space. The pre-trained convolutional encoder extracts features from the two-dimensional description of physical fields and adds them to the processed satellite feature representation through a fully connected neural network. From all the obtained variables, a unified vector is constructed and fed into the final fully connected neural network, which returns the reconstructed value of the sea surface salinity. Artificial neural network models were created using PyTorch framework [17].

For training all the considered models in this work, the Mean Squared Error (MSE) between the algorithm's output and the measured SSS values is taken as the loss function.

Table 1. Results from the ML Models Used

Model	RMSE, psu		
	"single-level"	"two-level"	total
Random Forest	4.39 ± 1.29	3.95 ± 0.98	2.64 ± 0.32
	3.43 ± 0.50	3.34 ± 0.52	
	2.07 ± 0.30	2.04 ± 0.29	
Gradient Boosting	3.75 ± 0.87	3.44 ± 0.86	2.16 ± 0.18
	2.70 ± 0.29	3.67 ± 0.66	
	1.71 ± 0.17	1.85 ± 0.28	
MLP	4.75 ± 0.78	4.56 ± 0.79	3.21 ± 0.29
	3.81 ± 0.56	5.33 ± 1.03	
	2.13 ± 0.34	2.32 ± 0.31	
MLP with ERA-5	4.25 ± 0.37		2.26 ± 0.26
	3.55 ± 1.07		
	1.83 ± 0.16		
SMAP SSS	6.06		3.15
	3.41		
	2.41		

4. RESULTS & DISCUSSION

4.1. SMAP SSS Accuracy

The figure demonstrates the distribution of satellite salinity errors relative to the measured values. The Root Mean Square Error (RMSE) is 3.15 psu, and the correlation coefficient (r) between them is 0.82. Due to the complex distribution of the target variable in machine learning algorithms (as shown in Fig. 2) and the diverse range of applications for the resulting models, errors are considered for the designated classes of "low," "medium," and "high" salinity values in addition to the overall error distribution. The error distributions for these classes are presented in the figure. The RMSE values for the designated classes are 6.06, 3.41, and 2.41 psu, respectively (Fig. 4c–4e)). Satellite algorithms perform best for high salinity values because the characteristics of water with such salinity values are closest to those of the World Ocean, for which satellite algorithms are verified with high quality.

4.2. SMAP SSS Improvement

To estimate SSS, classical ML models and artificial neural networks were examined in this study. The results are presented in Table 1.

Classical machine learning models have shown an improvement in SSS estimation quality compared to the standard satellite algorithms. The two classical models investigated are Random Forest and Gradient Boosting, as described in Section 3.2. Both models were explored using both "single-level" and "two-level" approaches. The hyperparameters of the models were tuned using the optuna framework.

As a result of applying the classical machine learning models, the best performance was achieved using the Gradient Boosting model, both on the designated classes of salinity values and the overall distribution. The RMSE values, obtained from comparing the measured salinity values with the model predictions (as shown in the table above), showed improvement compared to the results from the standard SMAP algorithms. The correlation coefficient was increased from 0.82 to 0.90. The characteristic distributions of the salinity error obtained using the best algorithms are presented in the figure.

The best machine learning model improved the quality of the standard algorithms. The overall error for all the data points was reduced to 2.15 psu after correction, compared to the previous 3.15 psu. The

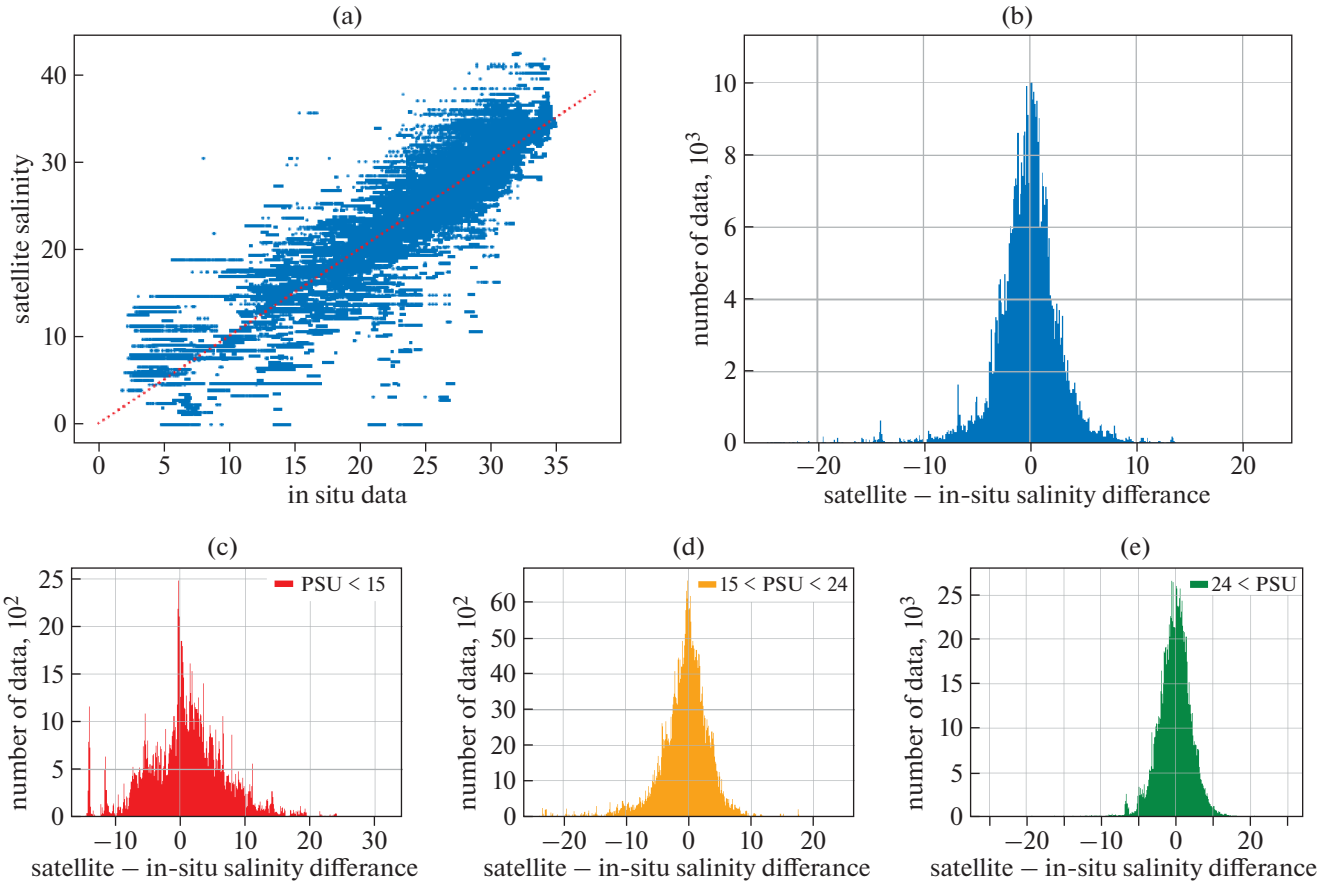

Fig. 4. Characteristics of SMAP sea surface salinity retrieval algorithm assessed using our dataset of measurements in the Arctic region: (a) a scatter diagram of approximated salinity compared to the measured SSS; and error distributions for (b) the full range of salinity, and also for (c) "low" , (d) "medium" and (e) "high" salinity values.

errors for the designated classes of "low", "medium," and "high" salinity values are 3.43 ± 0.86, 2.70 ± 0.29, and 1.71 ± 0.17 psu, respectively. The error distribution and the scatter plot for the predicted and the *in situ* values are presented in Fig. 5. Previously, these errors were 6.06 psu, 3.41 psu, and 2.41 psu, correspondingly. The highest quality, both for the standard algorithms and the models developed in this research, is achieved for "high" salinity values. The worst performance of the constructed algorithms is observed for "low" salinity values. This could be attributed to the relatively small amount of data in this range and the complex structure of stratified coastal waters.

Overall, based on the results of the classical models, we found that "single-level" models perform better on "high" and "medium" salinity values, while "two-level" models show higher quality on "low" salinity values.

In the neural network models, both "single-level" and "two-level" approaches were also considered. However, these models did not improve the quality compared to the classical approaches when applied to vector satellite data. Furthermore, the quality of some models was comparable to the initial quality of the standard satellite algorithms.

It is worth mentioning a neural network model that takes two-dimensional ERA-5 data as additional feature descriptors along with satellite data. The use of such a model is aimed to consider the climatic situation not only at the observation point but also in its vicinity. The results of this model are also presented in the table above, and its quality does not exceed the quality of the classical algorithms.

5. CONCLUSIONS & OUTLOOK

In this study, we presented an approach for enhancing the accuracy of sea surface salinity (SSS) retrieval in comparison with conventional SMAP satellite algorithms. The study incorporates various machine learning models, which include Random Forest, Gradient Boosting (Catboost), and artificial neural networks of various architectures, to investigate their

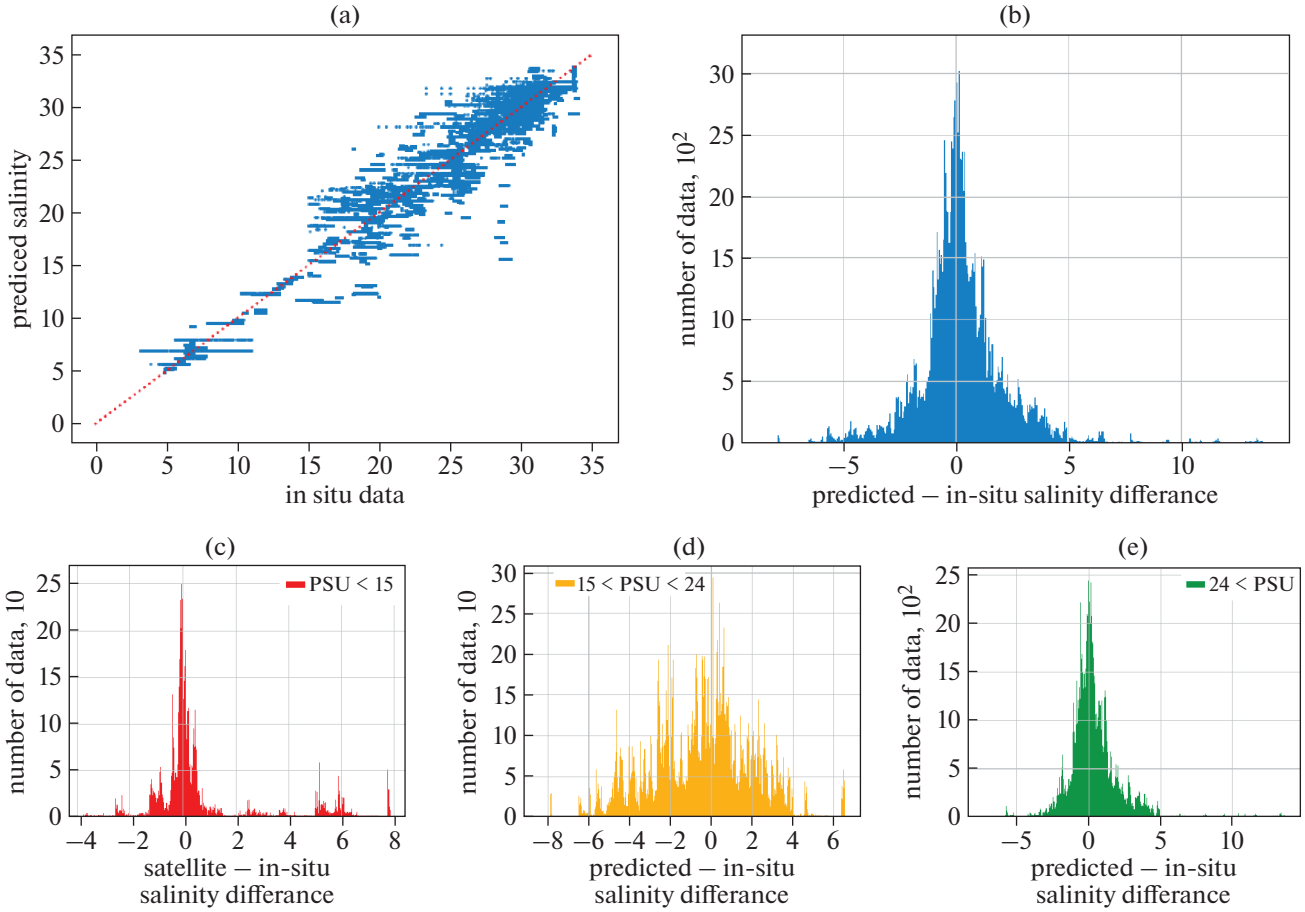

Fig. 5. Characteristics of a model delivering the best SSS approximation quality: (a) a scatter diagram for approximated salinity compared to the measured SSS, error distributions for (b) the full range of salinity and for (c) "low", (d) "medium" and (e) "high" salinity classes.

potential in SSS retrieval from remote sensing data. We explored the capabilities of these models comprehensively and found them improving the accuracy of SSS retrieval.

Classical methods, especially Gradient Boosting, demonstrated significantly higher quality of SSS retrieval compared to routine SMAP algorithms. Models trained on the entire dataset in general performed better on "high" and "medium" salinity values. At the same time, the composite models performing classification prior to the regression deliver higher results for "low" SSS.

The results of this study can be used to assess SSS values in the summer and autumn seasons in the Arctic Ocean for further scientific research.

FUNDING

This research was funded by the Moscow Institute of Physics and Technology Development Program (Priority-2030) (numerical modelling) and the Russian Science Foundation, research project 23-17-00087 (processing of in situ data).

CONFLICT OF INTEREST

The authors of this work declare that they have no conflicts of interest.

REFERENCES

1. P. J. Durack, T. Lee, N. T. Vinogradova, and D. Stammer, Nat. Clim. Change **6**, 228 (2016). https://doi.org/10.1038/nclimate2946
2. E. C. Carmack, Deep Sea Res., Part II **54**, 2578 (2007). https://doi.org/10.1016/j.dsr2.2007.08.018
3. E. P. Dinnat, D. M. Le Vine, J. Boutin, et al., Remote Sens. **11**, 750 (2019). https://doi.org/10.3390/rs11070750
4. N. Reul, S. A. Grodsky, M. Arias, et al., Remote Sens. Environ. **242**, 111769 (2020). https://doi.org/10.1016/j.rse.2020.111769
5. A. Supply, J. Boutin, J.-L. Vergely, et al., Remote Sens. Environ. **249**, 112027 (2020). https://doi.org/10.1016/j.rse.2020.112027
6. S. Qin, H. Wang, J. Zhu, et al., Acta Oceanol. Sin. **39**, 148 (2020). https://doi.org/10.1007/s13131-020-1533-0

7. W. Tang, S. Yueh, D. Yang, et al., Remote Sens. **10**, 869 (2018). https://doi.org/10.3390/rs10060869
8. E. C. Carmack, M. Yamamoto-Kawai, T. W. N. Haine, et al., J. Geophys. Res. Biogeosci. **121**, 675 (2016). https://doi.org/10.1002/2015JG003140
9. A. Matsuoka, M. Babin, and E. C. Devred, Remote Sens. Environ. **184**, 124 (2016). https://doi.org/10.1016/j.rse.2016.05.006
10. E. Jang, Y. J. Kim, J. Im, and Y.-G. Park, GIScience Remote Sens. **58**, 138 (2021). https://doi.org/10.1080/15481603.2021.1872228
11. D. Cho, C. Yoo, J. Im, et al., GIScience Remote Sens. **57**, 633 (2020). https://doi.org/10.1080/15481603.2020.1766768
12. T. D. Pham, K. Yoshino, and D. T. Bui, GIScience Remote Sens. **54**, 329 (2017). https://doi.org/10.1080/15481603.2016.1269869
13. L. Breiman, Mach. Learn. **45**, 5 (2001). https://doi.org/10.1023/A:1010933404324
14. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., J. Mach. Learn. Res. **12**, 2825 (2011).
15. A. V. Dorogush, V. Ershov, and A. Gulin, arXiv Preprint (2018). https://doi.org/10.48550/arXiv.1810.11363
16. T. Akiba, S. Sano, T. Yanase, et al., in *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, Anchorage, Alaska, 2019* (Association for Computing Machinery, New York, 2019), pp. 2623–2631 (2019). https://doi.org/10.1145/3292500.3330701
17. A. Paszke, S. Gross, F. Massa, et al., in *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing–NeurIPS Edition (EMC2-NIPS 2019)*, Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'AlchУй-Buc, E. Fox, and R. Garnett (IEEE, Vancouver, British Columbia, Canada, 2019), p. 8024.