Georgian Geographical Journal

Supervised Machine Learning in Drought Evaluation

Ana Palavandishvili[1,2]

[1] TSU, Vakhushti Bagrationi Institute of Geography, Tbilisi, Georgia

[2] Department of Engineering Physics/Informatics and Control Systems, GTU, Tbilisi, Georgia

* Corresponding author: ana.palavandishvili@tsu.ge

Georgian Geographical Journal, 2024, 4(2) 30-37

DOI:
https://journals.4science.ge/index.php/GGJ

Citation: Palavandishvili, A. Supervised Machine Learning in Drought Evaluation. *Georgian Geographical Journal* 2024, 4(2). 30-37
https://doi.org/10.52340/ggj.2024.04.02.04

Abstract

Among natural disasters, drought is one of the most common threats to many regions of the world and to Georgia as well. The monitoring and prediction methods of drought and precipitation distribution, the possibilities of their application in the reality of Georgia are considered in proposed work. Simulation methods such as Machine Learning (ML), namely Supervised Machine Learning (SML), optimal for similar complex tasks are presented as the alternative research methods. To conduct research, 1960-2022 period data were taken from database of National Environment Agency and the reanalysis data of the 1960-1990 Copernicus ERA5 rainfall, which were compared with the data of the stations on the territory of Georgia for the validation purpose. Standardized Precipitation Index (SPI) was selected as the research parameter. Using the prediction model and algorithm, drought-vulnerable areas in the Kakheti region were identified. As a result of comparison, lowest correlation rate was 0.309 at Shiraki, maximum was 0.657 at Omalo; minimum mean absolute error 1,662 at Udabno, the maximum 3,041 in Shilda. The smallest standard deviation 4,047 was fixes at Udabno, largest 7,624 at Lagodekhi. By analyzing stations data and satellite sources, it was determined that using the regression method of Machine Learning, it is sufficient to evaluate 1960-2000 period data for learning and 2001-2022 period data for training. The training time of Bagged Trees Optimized Algorithm was recorded as 326.21 sec, prediction speed ~ 7900obs/sec, RMSE - 0.5046, R2-0.64, MSE-0.25466, MAE-0.38065, training process minimum leaf size 19, and 40 iterations are assigned for optimization. CHIRPS satellite data were taken for next generation of the model. For prediction, it was necessary to calculate linear regression equation for each station. In the first case of forecast scenario, the amount of precipitation was determined from 0 cm to 10 cm. Gurjaani was highlighted, where forecast assumed SPI value from -0.008 to -0.901, and Kvareli- the SPI value from -0.002 to -0.138. The use of the presented ML model and algorithm for the analysis of precipitation distribution, drought monitoring and prediction is appropriate for Kakheti and other regions too in conditions of proper observation data base (60 years). It is recommended to use obtained results in early warning system for drought monitoring.

Keywords: Drought, Machine Learning, Big Data, early warning system.

Introduction

Big Data is a rapidly generated amount of information from a variety of sources and in a different format. Data analysis is the examination and transformation of raw data into interpretable information, while data science is a multidisciplinary field of various analyses, programming tools, and algorithms, forecasting analysis statistics, as well as machine learning that aim to recognize and extract patterns in raw data. The applicability of big data techniques is also significantly enhanced by the novel tools that support data collection and integration. The interoperability of the systems can be improved by data warehouses and the related ETL (extract, transform, and load) functionalities that can also be used to gather information from multiple models and data sources. Artificial intelligence (AI) and machine learning (ML) are also the key enabler technologies of big data analysis (Tatishvili et al., 2022a). Analysis of Big Data combines traditional methods of statistical analysis with computational approaches. The analysis of big data is a synthesis of quantitative and qualitative analyses. Climate computing combines multidisciplinary research regarding climatic, data, and system sciences to efficiently capture and analyse climate-related big data as well as to support socio-environmental efforts (Tatishvili et al., 2022). The significance of big data in climate-related studies is greatly recognized, and its techniques are widely used to observe and monitor changes on a global scale. It facilitates understanding and forecasting to support adaptive decision-making as well as optimize models and structures.

Drought is a climatic event that cannot be prevented, but interventions and preparedness to drought can help to: be better prepared to cope with drought; develop more resilient ecosystems; improve resilience to recover from drought; and mitigate the impacts of droughts.

Methods and Materials

Drought indices have been developed in large numbers and are widely used in drought evaluation, monitoring, and forecasting. Machine learning is a subset of artificial intelligence. Machine learning (ML) algorithms are a set of commands that allow systems to learn and improve from prior data without requiring complex programming. ML techniques have been used to implement prediction or forecasting of drought. These algorithms work by simulating a model from input datasets known as test sets and then using the model findings to forecast, predict, or make various types of judgments in various application domains. Various machine learning techniques are extensively used in the prediction of drought.

In order to conduct research, available data from the NEA database is used. Unfortunately, databases aren't perfect: data series aren't continuous and are less reliable, which is a big problem for index calculation and machine learning. Therefore, ERA5 reanalysis data were selected for historical data. Based on the comparison of satellite CHIPRS and IMERG information, it was determined that CHIRPS has a good correlation with the data of the stations in the territory of Georgia (Tatishvili et al., 2023). Those stations where 50% of the data were missing or did not correspond to reality were not subjected to analysis. Als, the satellite cannot perceive the mountainous region; the sea is not subject to observation because a filter (mask) has been applied, so in many cases it is impossible to obtain satellite data on the coastline; in this case, a nearby grid box of another grid must be selected.

Figure 1 shows a list of stations with insufficient data and whose data were not subjected to statistical analysis, and Figure 2 shows the results of the inventory of all (50) stations made using the program R-instat and covers the 2000-2020 period.

Figure 1 shows a list of stations with insufficient data and whose data were not subjected to statistical analysis

A comparison of CHIRPS and station data was made, for which the systematic error (BIAS) was calculated, which refers to the estimation of the difference between the monthly totals of precipitation measured from the satellite and the ground station.

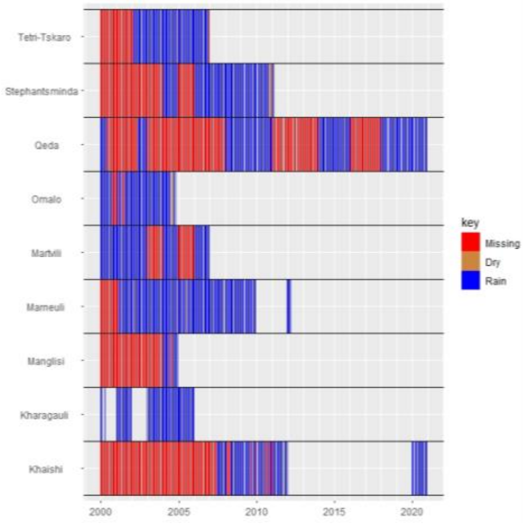

Figure 1. 2000 - 2020 list of stations that were not subjected to statistical analysis.

Results

Figure 2 shows the results of the inventory of all (50) stations made using the program R-instat and covers the period 2000-2020. The SPEI index was calculated for 1960 – 2022 period using the software package R—programming language Climpact2. Because data on daily precipitation and maximum and minimum temperature are needed to calculate this index, sufficient data for calculations were found for only 4 stations out of 17. As for the calculation of SPI, the total precipitation data of each month is required; the data of 10 stations that have at least 40 years of data series have been subjected to the analysis.

When calculating the three-month SPI, the index for the first January and February of the time series does not exist; to recover it, it is necessary to take the average of the month of January or February of each year of the entire series (Tatishvili et al., 2022b).

Correlation for the observations entire period, for all months and years, mean absolute errors and standard deviation were also calculated according to the same principle. The counting results are presented in Table 1.

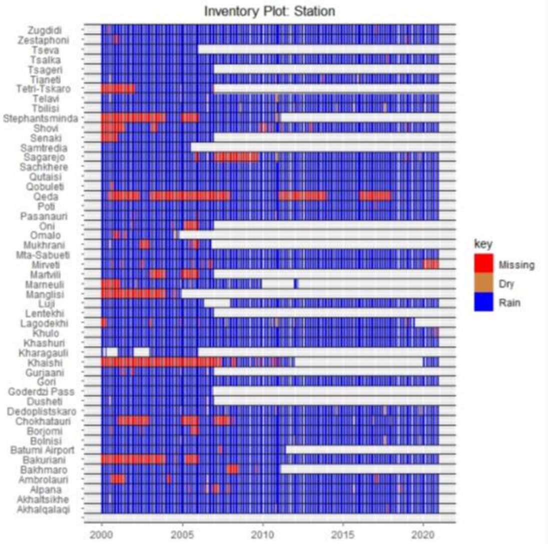

Figure 2. Inventory of 50 stations, the absence of data is marked in red, brown - dry period, blue - precipitation

Table 1. Calculated statistical parameters

Station	Correlation	Mean absolute error	Standard deviation	Correlation SPI3month_CHIRPS	Correlation IMERG SPI_3month	Correlation SPI3 month _CHIRPS check
Akhalqalaqi	0.66759	28.7	26.84318	0.416248	0.291970887	0.443102744
Akhaltsikhe	0.460294	38.1	33.55621	0.352573	0.484119499	0.352572913
Batumi Airport	0.58213	83.2	112.5929	0.448736	0.340454562	0.448736396
Borjomi	0.641079	24.1	25.26509	0.244318	0.310999227	0.244318316
Chokhatauri	0.578691	46.6	66.39357	-0.07615	0.165519382	-0.076151789
Gori	0.600696	18.8	23.40047	0.374053	0.424890006	0.374053213
Khashuri	0.602859	20.7	25.35372	0.384062	0.356057737	0.384061993
Manglisi	-	-	-	-		-
Mta-Sabueti	0.331451	41.3	51.98547	0.300716	0.31919611	0.300716118
Mukhrani	0.670568	22.5	24.8989	0.235464	-0.102300227	0.235464432
Pasanauri	0.633913	33.4	42.5678	0.290633	0.309023193	0.290632874
Poti	0.456032	72.8	111.9462	0.344086	0.375124979	0.344085861
Qeda	-	-	-	-		-
Qobuleti	0.623388	77.5	102.4052	0.286832	0.363684022	0.286831685
Qutaisi	0.66483	35.2	45.62226	0.483474	0.201479561	0.483474128
Sachkhere	0.592227	26.6	34.78904	0.27458	0.441969055	0.27458026
Sagarejo	0.619239	27.7	37.34099	0.076687	0.233896552	0.076687323
Senaki	0.647958	47.3	60.19786	0.747709	0.544157721	0.747709077
Shovi	0.717298	30.0	36.37001	0.440142	0.34740251	0.440141859
Stephantsminda	-	-	-	-		-
Tbilisi	0.678032	21.1	28.35372	0.208955	0.236983184	0.208954519
Telavi	0.693299	26.0	35.45118	0.271435	0.362963042	0.271434581
Tianeti	0.588555	22.0	31.88968	0.191648	0.321092386	0.191647776

Tsalka	0.504015	32.4	39.38712	0.285891	0.353743633	0.285890812
Zugdidi	0.603897	51.9	62.13524	0.477354	0.417972195	0.477354223

The calculations showed that in those stations where more than 50% of the data from the observation period are missing, for example, in Manglis, it is impossible to conduct a statistical analysis. A total of nine such stations were identified. From the analysis of the calculation of statistical characteristics, it is clear that the lowest correlation values are 0.33 at the Mta-Sabueti station and the maximum is 0.72 at the Shovi station; the minimum average absolute error is 18.8 at the Gori station and the maximum - 83.2 at the Batumi station; the smallest standard deviation is 23.4 at the Gori station; and the largest - 118.46 at Mirveti station, which is understandable considering the coastal and mountainous terrain. A high correlation coefficient better expresses the agreement between satellite and station data.

R - instat software was used to calculate Pearson's correlation and other statistical parameters. Stations that were omitted could not be analyzed.

Historical precipitation data from 1961–1991 were compared with ERA5 precipitation reanalysis data. 19 stations were selected for the Kakheti region. Of course, those points where there was insufficient data were not subjected to the analysis (Tatishvili et al., 2022c).

In the result of the data comparison, the lowest correlation index (0.309) was revealed at the Shiraki station, the maximum (0.657) at the Omalo station. The minimum average absolute error (1.662) was recorded at Udabno station, the maximum (3.041) at Shielda station. The smallest standard deviation (4.047) is at Udabno station, the largest (7.624) at Lagodekhi station. By comparing the historical and ERA5 reanalysis data of the station for 1961 – 1991 period, reanalysis data can be used as an alternative database for the model, both in Kakheti and in the entire territory of Georgia.

Table 2. Comparison of station and ERA5 reanalysis precipitation1961-1991 data

Stations	Standard deviation	Correlation	Mean abs. error
Akhmeta	5.76054943	0.5936142	2.22960019
Artana	6.81349292	0.58776933	2.34396992
Birkiani	7.62295759	0.63601722	2.50106601
Dedoplistskaro	5.177286	0.57295248	2.34935663
Gombori	5.48090201	0.57076494	2.1249113
Gurjaani	6.29609913	0.61040494	2.13628114
Kachreti	5.37111452	0.5144723	2.64170272
Khvareli	7.51038719	0.58870058	2.30064283
Lagodekhi	7.62411052	0.55838778	2.39522353
Lechuri	7.26083196	0.48469664	2.65489445
Omalo	5.06645709	0.65703807	2.28134972
Sabue	7.18089762	0.5644877	2.44340144
Sagarejo	6.09202268	0.60668778	2.295695
Shilda	6.35923532	0.48506866	3.04148853
Shiraki	5.45504798	0.30906783	2.58658036
Tsiauri	4.83284442	0.5361194	1.87135863
Telavi	6.16741437	0.61484533	2.13956842
Tsnori	5.19081531	0.62222577	1.96229826
Udabno	4.04739384	0.51336395	1.66217209

Discussions

At the initial stage of machine learning, the station coordinates, station name, year, month, monthly precipitation total, and 3-month precipitation index were included in the training (Mastering Machine Learning). A 30-year observation period was analysed at all. As a result of counting, we got a good result only for those stations that are located close to each other and the overfit for the stations located at a relatively long distance, e.g., Omalo, Akhmeta.

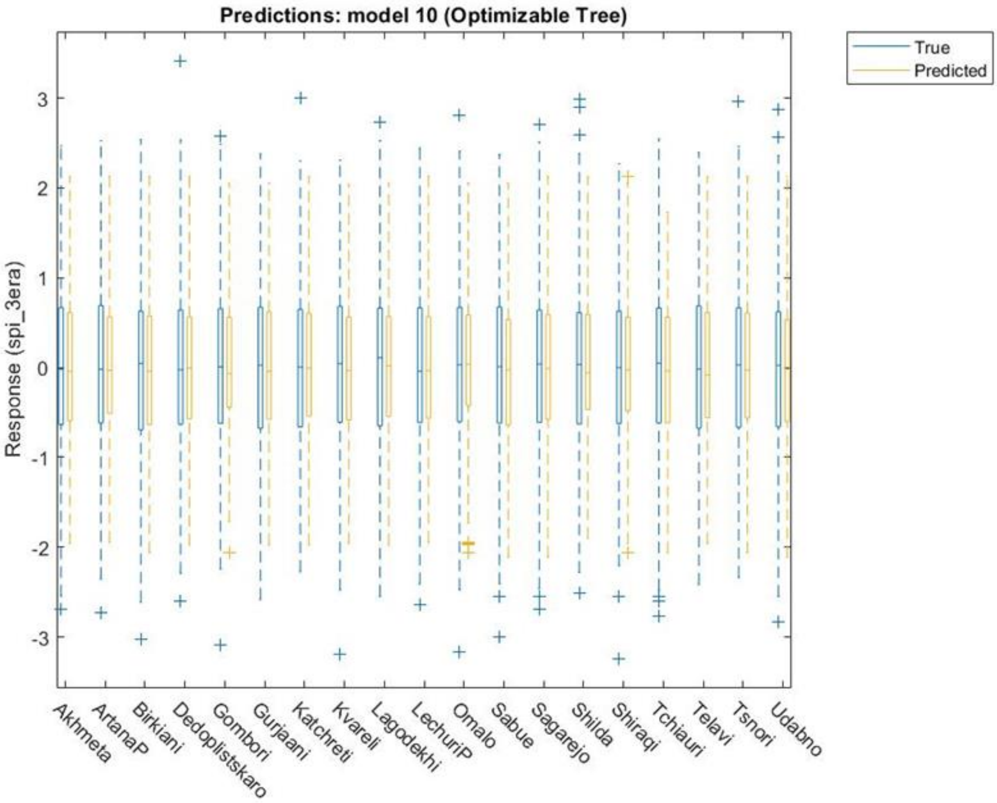

Figure 3. Comparison of the model created by the SVM optimized algorithm with the real data. True data values are marked in blue and values generated by the model in yellow

The parameters for estimating the model generated by the support vector machine (SVM) of the first model are: RMSE - 0.58848, R2 _0.64, MSE_0.34631, MAE_0.45479, prediction speed ~430000 obs/sec (observations per second), training time - 30,296 s, minimum leaf size - 45 (optimized parameter), hyperparametric search ranking: minimum leaf size_1-3420. All these parameters are needed to estimate the model. To avoid excessive adjustment, 30 additional stations in the Kartli region were added.

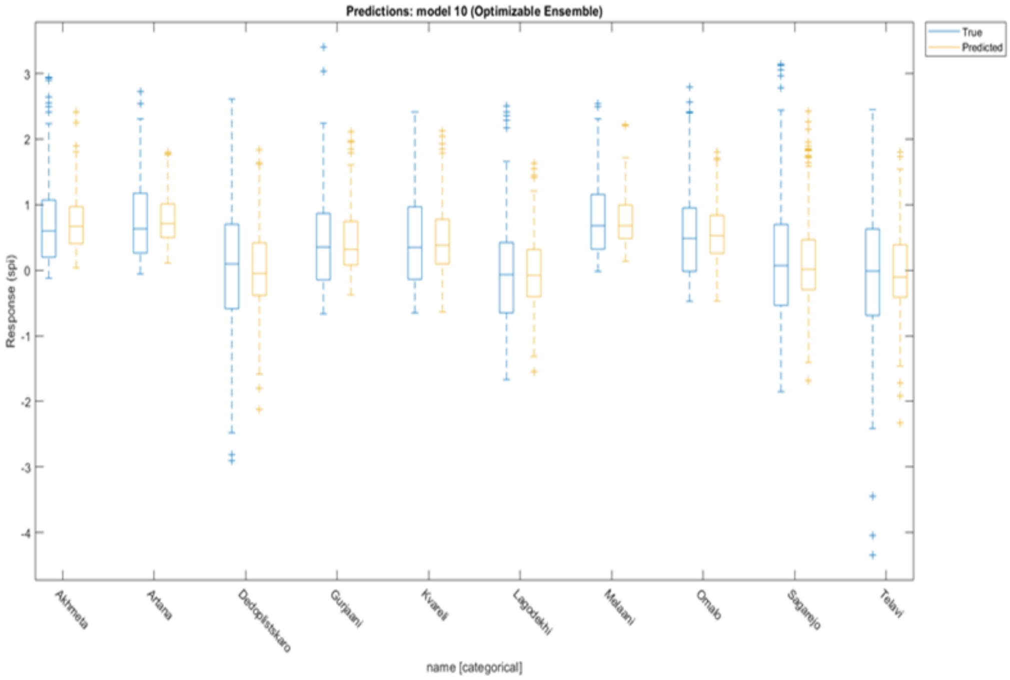

Figure 4. Comparison of the model generated by MatLab and the actual observation data

Figure 4 shows a visual comparison of the actual and predicted results generated by MatLab. Actual data value is marked in blue, predicted data value in yellow.

Bagged trees assembled by an optimized algorithm showed the best results in regression algorithm training. Learning time is 326.21 s, prediction speed - ~ 7900 observations/second, characteristic parameters: RMSE - 0.5046, R2 - 0.64, MSE - 0.25466, MAE - 0.38065.

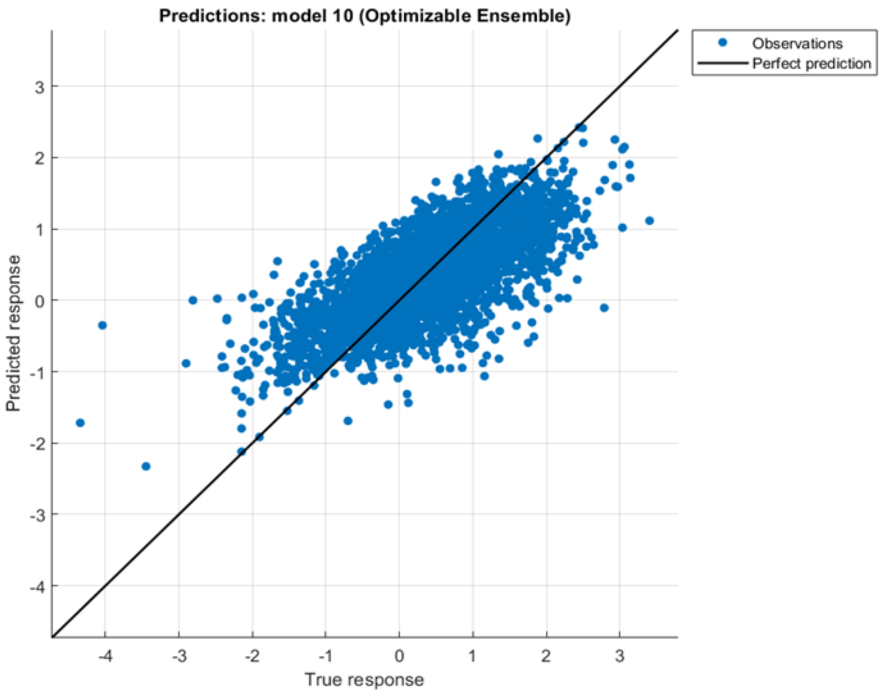

Figure 5. Prediction response generated by MatLab

Figure 5 shows the observation points and the best forecast. Based on the analysis of this drawing, we can judge what kind of model it is. If the data (points) are very scattered, this indicates underfitting, while points that are close to the prediction curve (line) indicate overfitting (McHanay, 2014). When many points are gathered in one area, it can be concluded that the model is a good fit, i.e., satisfactory.

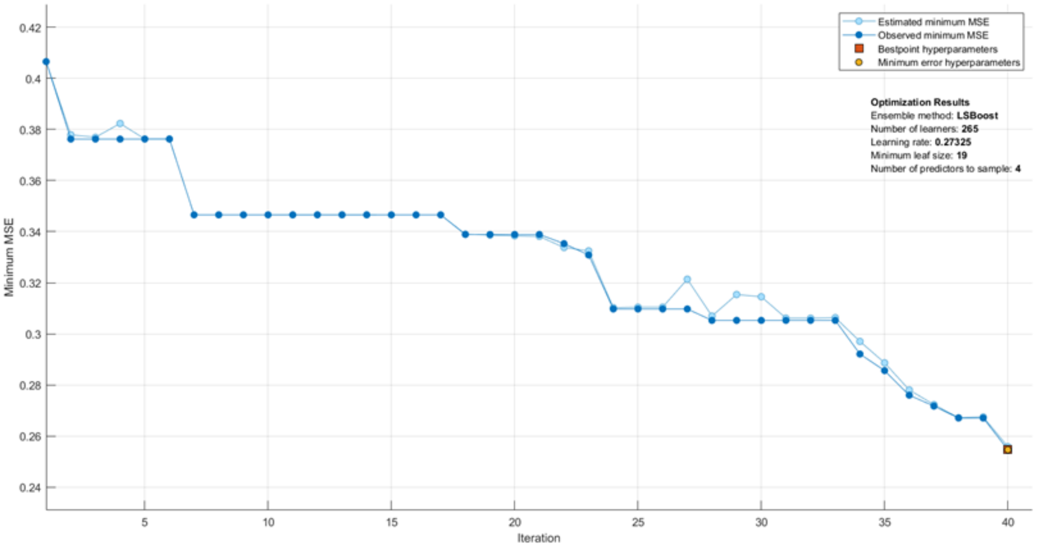

Figure 6. Learning process of optimized Bagged Trees generated by MatLab

Figure 6 depicts the optimized Bagged Trees learning process. The minimum leaf size is 19, with 40 iterations instead of 30 iterations assigned for optimization. Estimated minimum mean square deviation is marked in blue, observed minimum mean square deviation in blue. The learning rate is 0.27325.

After the training stage, the model was subjected to fitting to the data of 2001-2022 for further processing. It is also worth noting that for the adjustment and forecasting process, those stations that operated between 2001 and 2022, at least for several years, were selected. They were then filled with CHIRPS satellite data (Palavandishvili, 2021).

To perform the prediction, it was necessary to calculate a linear regression equation for each station using the polyfit function in MatLab space. After calculating the coefficients, polynomials were calculated in the MatLab space with the polyval function.

$$y=ax+b$$

From the parameters included in this equation, the following were selected: a - the standardized precipitation index calculated by the model, x - the value of the sum of the monthly precipitation for a specific scenario, b - the standardized value of the actual precipitation.

In the first case of the forecast scenario, the amount of precipitation was determined from 0 cm to 10 cm because we are interested in the process of drought development in the case of a small amount of precipitation.

Table 3. Standardized Precipitation Index (SPI), 0 cm to 10 cm inclusive precipitation forecast for some regions of Kakheti

Precipitation forecast (sm)	Dedoplistskaro	Sagarejo	Gurjaani	Khvareli	Lagodekhi	Omalo	Telai
0	-0.270	-0.336	-0.008	- 0.002	-0.524	0.270	-0.568
1	-0.021	-0.057	-0.097	-0.012	-0.269	0.429	-0.325
2	0.227	0.220	-0.186	-0.026	-0.014	0.587	-0.083
3	0.475	0.498	-0.275	-0.041	0.240	0.746	0.159
4	0.724	0.776	-0.365	-0.054	0.495	0.905	0.402
5	0.973	1.054	-0.454	-0.068	0.750	1.064	0.645
6	1.222	1.333	-0.543	-0.082	1.005	1.222	0.888
7	1.470	1.611	-0.633	-0.096	1.260	1.381	1.131
8	1.719	1.889	-0.722	-0.110	1.514	1.540	1.374
9	1.968	2.167	-0.811	-0.124	1.769	1.698	1.617
10	2.217	2.445	-0.901	-0.138	2.024	1.857	1.860

Conclusion

The analysis showed that Gurjaan and Kvareli are more vulnerable to a small amount of precipitation, and less vulnerable regions were also identified: Dedoplistskaro, Sagarejo, and Lagodekhi. When analyzing these results, station altitude and humidity should be considered, as well as the fact that in the case of Artana and Melaani stations some satellite data were not found; that's why the model could not cover the full Kakheti region.

To increase the accuracy of such research, it is necessary to determine additional parameters—temperature, humidity, soil moisture, and indices calculated from the satellite—their validation and adjustment for the machine learning algorithm. Also, the period of education and training should be determined separately; the appropriate method for forecasting will be selected, by means of which we will get the regression equation.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

A. Palavandishvili conceived of the presented idea, performed the analytic calculations, and took the lead in writing the manuscript.

ORCID iD

Ana Palavandishvili https://orcid.org/0000-0002-7254-685X

Reference

Begalishvili, N., Robitashvili, G., & Tatishvili, M. (2005). The Investigation of Precipitation Efficiency of Convective Clouds. *Bulletin of the Georgian Academy of Science*.

Guidelines on the Definition and Characterization of Extreme Weather and Climate Events, World Meteorological Organization (WMO). 2023, 36 pp., ISBN 978-92-63-11310-8. https://library.wmo.int/doc_num.php?explnum_id=11535

Mastering Machine Learning, A Step-by-Step Guide with MATLAB, Mathwork production. 22 pp., Mastering Machine Learning: A Step-by-Step Guide with MATLAB - MATLAB & Simulink (mathworks.com)

Mathwork - Matlab Statistics and Machine Learning Toolbox documentation (2016). stats.pdf (mathworks.com)

Mathwork - Statistics and Machine Learning Toolbox Release notes rn.pdf (mathworks.com)

Mastering Machine Learning, A Step-by-Step Guide with MATLAB, Mathwork production. 22 pp., Mastering Machine Learning: A Step-by-Step Guide with MATLAB - MATLAB & Simulink (mathworks.com)

McHanay, R. (2014). Understanding computer simulation, bookboon, eBook company.

Palavandishvili., A. (2021). Structural data set in environmental issued. The Regional Student Scientific and Practical Conference Digital Transformation in Engineering Human-Computer Interaction, *Georgian Technical University (GTU)*, Faculty of Informatics and Control Systems.

Tatishvili, M., Palavandishvili, A., Tsitsagi, M., & Suknidze N. (2023). The Big Data for Drought Monitoring in Georgia. *Springer, Cham*. 131-142

Tatishvili, M., Palavandishvili, A., Tsitsagi, M., & Suknidze, N. (2022a). The use of structured data for drought evaluation in Georgia. *Journal of the Georgian Geophysical Society, Physics of Solid Earth, Atmosphere, Ocean and Space Plasma*, *25(1)*, 45-51

Tatishvili, M., Palavandishvili, A., Tsitsagi, Gulashvili, Z., M., & Suknidze, N. (2022b). Drought Evaluation Based on SPEI, SPI Indices for Georgian Territory. International Conference of Young Scientists *“Modern Problems of Earth Sciences”. Proceedings, Publish House of Iv. Javakhishvili Tbilisi State University*, Tbilisi, November 21-22, 119-121.

Tatishvili, M., Palavandishvili, A., & Samkharadze, I. (2022c). Disaster Risk Reduction and Climate Resilience in Nature Based Solutions Using In-Situ and Satellite data for Georgia Sustainable Development. *International Conference of Young Scientists “Modern Problems of Earth Sciences”. Proceedings*, Publish House of Iv. Javakhishvili Tbilisi State University, 116-118.

Tatishvili, M., Palavandishvili, A., & Samkharadze, I. (2022). Disaster Risk Reduction and Climate Resilience in Nature Based Solutions Using In-Situ and Satellite data for Georgia Sustainable Development. *International Conference of Young Scientists “Modern Problems of Earth Sciences”*. Publish House of Iv. Javakhishvili Tbilisi State University. 116-118.

Tatishvili, M., Megrelidze, L., & Palavandishvili, A. (2021). Study of the mean and extreme values, intensity, and recurrence variability of meteorological elements based on the 1956-2015 observation data. *Journal of the Georgian Geophysical Society, Physics of Solid Earth, Atmosphere, Ocean, and Space Plasma*, 24. 73-77.