Machine learning methods' comparison for land surface temperatures forecasting due to climate classification*

Tetiana Hovorushchenko[1,†], Vitalii Alekseiko[1,*,†] and Vitaly Levashenko[2,†]

[1] *Khmelnytskyi National University, Institutska str., 11, Khmelnytskyi, 29016, Ukraine*
[2] *Zilina University, Univerzitná 8215, 010 26 Žilina, Slovakia*

Abstract
The application of machine learning methods for short- and medium-term forecasting of the average monthly temperature of the Earth's surface, taking into account climatic zoning, is considered. The peculiarities of predicting the temperature of a moving surface in the context of the regression problem using machine learning methods are described. A comparison of the forecasting accuracy of methods based on metrics was made. Peculiarities of calculating the metrics, according to the values of the investigated parameters, are considered. The speed of operation of the methods was analyzed and statistical indicators were calculated. To visualize the effectiveness of the methods, Taylor diagrams were constructed. The most effective methods for forecasting the temperature of the Earth's surface have been determined.

Keywords
machine learning (ML), forecasting, land surface temperature, climate zone, models' evaluation, regression, climate changes.

1. Introduction

The problem of changing the temperature of the Earth's surface has a wide range of consequences that affect almost all aspects of people's lives, including the spheres of agriculture, health care, economy, energy, infrastructure, tourism, forest and water management.

Social aspects are also particularly acute. Climate changes increasingly become the cause of population migration, climate refugees appear, which in turn leads to changes in the social structure of communities and creates new civilizational challenges. Today, climate change is felt all over the planet, but certain regions are particularly vulnerable [1, 2]. In view of this, the question arises of predicting possible climate changes in order to develop strategies for avoiding and mitigating the consequences.

Nowadays, there are several approaches to predict climate parameters, but the rapid development of machine learning technologies has led to the emergence of new and effective methods that are competitive to numerical knowledge-based alternatives [3, 4]. Machine learning (ML) technologies are capable of processing large volumes of data more efficiently and capturing patterns, which makes them extremely effective in solving the problem of forecasting.

2. Peculiarities of land surface temperature forecasting

Forecasting of climate parameters is extremely relevant in the context of modern climate changes [5]. One of the main and most widely studied parameters is temperature. Modern scientific research is aimed at determining the main trends in air [6, 7], land [8] and water [9] surface temperature

AdvAIT-2024: 1st International Workshop on Advanced Applied Information Technologies, December 5, 2024, Khmelnytskyi, Ukraine - Zilina, Slovakia
* Corresponding author.
† These authors contributed equally.
tat_yana@ukr.net (T. Hovorushchenko); vitalii.alekseiko@gmail.com (V. Alekseiko); vitaly.levashenko@fri.uniza.sk (V. Levashenko)
0000-0002-7942-1857 (T. Hovorushchenko); 0000-0003-1562-9154 (V. Alekseiko); 0000-0003-1932-3603 (V. Levashenko)

changes. The study of these parameters has a key influence on determining the priorities of greening, urban planning and landscape design [10].

Forecasting the temperature of the Earth's surface is associated with some complexities and peculiarities due to the dynamic nature of the climate system and the Earth's surface itself. The key factors that determine these features are:

– spatial variability;
– temporal variability;
– feedback mechanisms;
– uncertainties in models;
– anthropogenic impact;
– extreme events;
– availability and quality of data.

The Earth's surface temperature varies greatly among regions due to factors such as latitude, proximity to oceans, elevation above sea level, and types of land cover (such as forests, deserts). Forecasting must take into account these spatial variations, which may affect local weather conditions.

It should be noted that the temperature of the Earth's surface fluctuates not only seasonally, but also daily due to diurnal cycles (day-night). However, when studying the main trends in the average monthly temperature, it is advisable to ignore daily cycles. Also, weather patterns and climate phenomena such as El Niño and La Niña can cause interannual variability. In addition, they are able to influence long-term temperature trends.

The Earth's climate system is driven by various feedback mechanisms, such as the albedo effect (the reflectivity of the Earth's surface), the concentration of greenhouse gases (e.g. CO_2) and the accumulation of heat in the ocean. These feedbacks can amplify or weaken temperature changes, making predictions difficult.

Climate models that simulate Earth's climate include physical processes such as radiation, convection, and ocean currents. These models contain uncertainties due to imperfect knowledge of the parameters and the complexity of the interaction between different components of the climate system.

Another important factor is human activity, including industrial emissions, land-use change, and urbanization [11], which contribute to warming trends [12]. Predicting how these factors will evolve and interact with natural climate variability adds another layer of complexity to temperature projections.

Forecasting extreme temperature events, such as heat waves and cold snaps, requires understanding not only trends in average temperature, but also the likelihood and intensity of such events under changing climate conditions.

Surface temperature forecasting is based on historical data from weather stations, satellites, and other sources. Ensuring the accuracy and reliability of this data, especially in remote or data-poor regions, can present challenges for forecasting models.

Solving these complexities involves integrating observations, improving modeling methods, and understanding of the Earth's climate system. Technological advances and computing power continue to improve our ability to more accurately predict surface temperatures on different time scales.

3. Methodology

3.1. Dataset

In the research it was used a dataset GlobalLandTemperatures [13] with Creative Commons License (CC0: Public Domain) from Kaggle.

This dataset includes Earth's surface temperatures data from 1743 to 2013. Original tables content following information:

– dt (includes month and year, when the temperature was observed);
– AverageTemperature (average monthly temperature);
– AverageTemperatureUncertainty (with uncertainty values of measurement);
– Country (includes country or territory, where the temperature was observed);

Due to needs of research the dataset was modified. It was added columns 'ClimateZone' with abbreviation of climate zones according to the World Climate Data [14] and 'MainClimateZone' with letter, which means belonging to one of five main climate zones. Table 1 shows number of countries of each main climate zone.

Although the dataset cannot be fully called balanced, this is explained by the peculiarities of the location of countries on the globe and the geopolitical situation. There are some important aspects:

– Area of Countries;
– Geopolitical factors;
– Selecting Data Sources.

First of all, large countries can have a variety of climate zones. For example, some countries cover a vast territory with varying climate conditions: from arctic to temperate in Canada or from temperate to arid in the USA. This may result in uneven presentation of temperature data.

Secondly, political, economic and sociocultural differences between countries can also affect the balance of the dataset. For example, access to climate observation technologies may be uneven across countries, which may affect the accuracy of the data.

Finally, different countries may have different climate monitoring systems and different data sources. Some countries may be active in collecting data, while others may be less active. This can also affect the balance of the dataset.

Table 1
Number of countries of each main climate zone

Main Climate Zone	Number of countries
A	94
B	45
C	64
D	26
E	6

In general, the imbalance of the dataset with the temperatures of the earth's surface by country is a complex issue associated with many factors. For more accurate climate analysis and modeling, it is important to consider all these aspects.

In the research was used data with similar values of uncertainty, but sometimes this values are different, so forecasting may be more or less accurate for some regions. To avoid any discrimination, it was used data for all countries and territories with relevant information.

3.2. Machine learning methods

It was conducted a study of the operation of various methods for different climate zones. To do this, it was developed several models to forecast temperature for the period from 2000 to 2013. Temperature data up to the year 2000 were used for model fitting.

The following methods were chosen for the study:

– neural network;
– decision trees;

– random forest;
– K nearest neighbors;
– method of support vectors;
– gradient boosting;
– Ada boost;
– XG boost;
– light GBM.

Due to the climatic features of different regions of the Earth, it is advisable to conduct separate studies for each of the climatic zones in order to identify the methods that are best adapted to the corresponding temperature dependencies [15, 16].

3.2.1. Neural Network

A neural network (NN) is a set of algorithms modeled after the human brain designed for pattern recognition. A neural network interprets the data using a kind of machine perception, labeling or clustering of the raw data. Neural networks consist of layers of interconnected nodes ("neurons") that process input data, learn from it, and make decisions based on learned patterns. Each node is assigned a weight that is adjusted during learning to minimize the prediction error.

In the context of regression tasks for predicting numerical series, neural networks can model complex relationships between inputs and outputs. Recurrent neural networks (RNNs), long-short-term memory (LSTM) networks, and supervised recurrent units (GRUs) are particularly well-suited to time series forecasting because they can capture temporal dependencies in data. By learning from historical data, these networks learn patterns and trends that can be used to predict future values.

3.2.2. Decision Trees

Decision Tree (DT) is a non-parametric supervised learning method used for classification and regression problems. It partitions the dataset into subsets based on the most important feature at each node, making decisions based on feature values. Each branch of the tree represents a decision rule, and each leaf represents an outcome. The decision rule can be represented as:

$$\text{if } x_i < a \text{ then go to left subtree, else go to right subtree,} \quad (1)$$

where:
x_i – feature;
a – threshold.

In the context of a regression problem, decision trees can be used by partitioning the data into subsets based on input feature values and predicting a numerical value for each subset [17]. In time series forecasting, decision trees can model the relationship between time-based features and a target variable. Although decision trees are easy to understand and interpret, individual decision trees can be prone to overfitting and as a result may not perform well with complex patterns, but at the same time the method is fundamental to ensemble methods such as random forest and gradient boosting.

3.2.3. Random Forest

Random Forest (RF) is an ensemble learning method. The work of the method is based on building several decision trees during the learning process and deriving class membership (classification task) or average prediction (regression task) of individual trees [18]. A random forest combines the simplicity of decision trees with improved accuracy, robustness, and robustness to overfitting by averaging the results of multiple trees that may individually be subject to overfitting [19].

$$y = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad (2)$$

where:

y_i – prediction of the i-th tree;

N – total number of trees.

In the context of numerical series prediction, each tree is trained on a random subset of data and features, and their predictions are averaged to produce a final prediction. Random forests are quite robust and handle a large number of input variables well.

3.2.4. K-nearest neighbors

K-Nearest Neighbors (KNN) is a simple instance-based learning algorithm that classifies a data point based on how its neighbors are classed. In KNN, the parameter "K" represents the number of nearest neighbors to consider. The algorithm calculates the distance between the new data point and the training points and then assigns a class based on the K-nearest neighbor majority votes.

KNN can be applied to regression tasks by averaging the numerical values of the K-nearest neighbors. For time series forecasting, KNN can predict the future value by finding similar historical patterns and averaging their subsequent values. KNN is simple to implement, but can be computationally expensive and sensitive to the choice of K as well as the distance metric used.

3.2.5. Support Vector Regression

Support Vector Machine (SVM) is a supervised learning algorithm that can be used for classification or regression. The algorithm of the method consists in finding the hyperplane that best divides the data into classes [18]. In cases where the data cannot be partitioned linearly, SVM uses a transformation of the data into a higher dimensional space where a hyperplane can be used for partitioning.

In regression problems, the support vector method is known as support vector regression (SVR). SVR tries to find a function that deviates from the actual observed values by an amount that does not exceed a given threshold and is as smooth as possible. For time series forecasting, SVR can capture the underlying trend and seasonality in the data, although this often requires careful parameter tuning and kernel selection.

3.2.6. Gradient Boosting

Gradient Boosting (GB) is a complex technique that sequentially builds models, where each new model tries to correct mistakes made by previous models. This approach uses a gradient descent algorithm to minimize the loss function. The method is powerful for both classification and regression tasks [19]. It is very efficient and accurate in forecasting, although it may require significant resources for intensive calculations.

The loss function in regression problems is often represented by mean squared error or mean absolute error. Variations of the gradient boosting method, in particular XGBoost and LightGBM, are known for their high accuracy and ability to handle complex datasets.

3.2.7. Ada boost.R

Adaptive Boosting (AB, AdaBoost) is an ensemble learning technique that combines several weak classifiers to create a strong classifier. This method focuses on cases that previous classifiers misclassified and adjusts their weights accordingly, thus increasing the accuracy of the model. Each subsequent model in the sequence is tuned to correct the errors of the previous ones, making it highly effective at improving forecasting performance.

AdaBoost can be adapted for the regression problem (AdaBoost.R). In this context, the method combines the predictions of several weak methods, typically decision trees, to create a strong predictive model. Each such method focuses on correcting the mistakes of the previous ones. For numerical series prediction, AdaBoost.R can improve prediction accuracy by highlighting hard-to-predict data points during fitting.

3.2.8. XG boost

XGBoost (XGB – Extreme Gradient Boosting) is a powerful and efficient implementation of gradient boosting. The method includes numerous optimizations such as parallel processing, tree pruning, and missing value handling, making it faster and more accurate than traditional gradient boosting methods [16]. XGBoost is widely used in competitive machine learning due to its performance and flexibility.

XGBoost is a powerful tool for regression tasks and is widely used for predicting numerical series. It includes optimizations such as parallel processing and regularization to prevent overfitting. XGBoost builds trees sequentially, where each tree aims to reduce the residual errors of previous trees. The method is known for its high performance and scalability, making it a popular choice for forecasting tasks.

3.2.9. Light GBM

LightGBM (LGBM – Light Gradient Boosting Machine) is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be highly efficient and scalable, suitable for large datasets. LightGBM uses histogram-based algorithms, which provides faster fitting and less memory usage compared to traditional gradient boosting frameworks.

LightGBM is particularly effective for regression tasks, including predicting number series. The method uses histogram-based algorithms to efficiently group and divide data. LightGBM handles large datasets and high-dimensional data efficiently, making it suitable for numerical series prediction. It builds trees sequentially, with each tree correcting the errors of previous ones, similar to other gradient boosting methods.

3.3. Models' evaluation

To evaluate the effectiveness of the predictive model in the regression problem, various aspects of performance are measured. The most common metrics include [20, 21, 22]:

– Mean Absolute Error (MAE):

Indicates the average of the absolute differences between predicted and actual values. Estimates the accuracy of forecasts without considering the direction of errors [22].

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{3}$$

where n – number of observations;

y_i – the actual value of the i-*th* observation;

$\hat{y}_i$ – the predicted value of the i-*th* observation.

– Mean Squared Error (MSE):

Indicates the mean of the squared differences between predicted and actual values, giving greater weight to larger errors.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

– Root Mean Square Error (RMSE):

The square root of the root mean square error has the same units as the raw data, making it easier to interpret [22].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{5}$$

– R-squared (R^2):

Indicates the proportion of variance in the dependent variable that can be predicted from the independent variable(s). The value ranges from 0 to 1, with higher values indicating a better fit of the model to the tasks at hand.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \quad (6)$$

where $\bar{y}_i$ – the average among the actual values.

– Mean Absolute Percentage Error (MAPE):

Determines the average value of the absolute percentage errors. MAPE expresses the error as a percentage of the actual values.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (7)$$

– Symmetric Mean Absolute Percentage Error (sMAPE):

It is a type of MAPE that takes into account positive and negative deviations symmetrically.

$$sMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \quad (8)$$

– Mean Bias Deviation (MBD):

Determines the average bias in the forecasts, indicating whether the model is systematically over- or under-predicting.

$$MBD = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (9)$$

– Median Absolute Error (MedAE):

Is the median of the absolute differences between the predicted and actual values. The mean absolute error is less sensitive to outliers compared to the mean absolute error (MAE), making it a reliable measure of model performance when there is significant outliers in the data.

$$MedAE = median(|y_i - \hat{y}_i|) \quad (10)$$

A Taylor diagram is a graphical tool used to evaluate the performance of models by comparing their results to observations. The chart combines three statistics into one graph: correlation coefficient, standard deviation, and root mean square error (RMSE).

The standard deviation σ represents the variability or spread of the data. Calculated for both observation data and model data:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}, \quad (11)$$

where:

n – number of observations;

x_i – each individual observation;

$\bar{x}$ – the mean value of the observations.

Correlation coefficient r between observation data and model data. It indicates how well the model results match the observed data in terms of patterns and time intervals.

$$r = \frac{\sum_{i=1}^{n}(x_{obs_i} - \bar{x}_{obs})(x_{model_i} - \bar{x}_{model})}{\sqrt{\sum_{i=1}^{n}(x_{obs_i} - \bar{x}_{obs})^2 \sum_{i=1}^{n}(x_{model_i} - \bar{x}_{model})^2}}, \quad (12)$$

where:

x_{obs_i} – individual values of observations;

x_{model_i} – individual model values;

$\bar{x}_{obs}$ – the mean value of the observations;

$\bar{x}_{model}$ – model mean value.

The centered root mean square error E' reflects the total difference between the model output and the observations, taking into account both the variance and the bias.

$$E' = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(x_{model_i} - \bar{x}_{model}) - (x_{obs_i} - \bar{x}_{obs})\right]^2} \quad (13)$$

In this way, a Taylor chart allows you to compare multiple models on a single graph, making it easier to visualize and interpret the relative performance of different machine learning techniques.

The chart layout makes it easy to see how close the model's performance is to ideal (represented by a control point where the correlation is 1, the standard deviation matches the observed, and the RMSE is zero).

Thus, the Taylor plot is a powerful tool for evaluating and comparing the performance of machine learning methods, providing a visual and quantitative assessment of their ability to accurately reproduce observed data patterns.

4. Results

In general, most machine learning methods have demonstrated high predictive accuracy on test data. The calculated metrics are presented in Tables 3.1 – 3.5, separately for each climate zone.

For the tropical climate zone (Table 2), KNN (72.1%) and LGBM (73.6%) methods showed the highest efficiency according to the R2 metric. At the same time, their MAPE was 0.68% and 0.65%, respectively. However, the complex temperature patterns associated with the geographical location of most countries do not allow a highly accurate forecast to be made. In particular, this is due to the location of the studied territories in different hemispheres, as well as some countries located in both hemispheres. The impossibility of making a high-precision forecast necessitates further studies of the tropical climate zone, in particular by conducting separate studies for different hemispheres, as well as climate subzones.

Table 2
ML methods' metrics evaluation for land surface temperature forecasting in zone A

Method	Metrics							
	MAE	MSE	RMSE	R2	MAPE, %	sMAPE, %	MBD	MedAE
NN	0.3721	0.2117	0.4601	0.0854	1.4097	1.4189	0.1909	0.3157
DT	0.3039	0.1418	0.3766	0.3872	1.1489	1.1586	0.2734	0.2622
RF	0.2508	0.1041	0.3226	0.5503	0.9475	0.9534	0.1811	0.2383
KNN	0.1803	0.0645	0.254	0.7212	0.6812	0.6841	0.0857	0.1422
SVR	0.6606	0.5412	0.7356	−1.338	2.4935	2.5322	0.6590	0.6462
GB	0.2643	0.1154	0.3397	0.5015	0.9963	1.0037	0.2232	0.2432
AB	1.0841	1.6211	1.2732	−6.003	4.1325	4.2554	1.0781	1.1265
LGBM	0.1708	0.0611	0.2471	0.7362	0.6471	0.6486	0.0238	0.1275
XGB	0.2964	0.1363	0.3692	0.4110	1.1201	1.1293	0.2647	0.2625

In the arid climate zone (Table 3), temperature patterns are clearly observed, which are common to both hemispheres, except for the shift caused by different seasons in different hemispheres. This allows for accurate forecasting using various methods. Thus, NN, DT, RF, KNN, GB, LGBM and XGB

show a high R^2 metric value (above 96%). In addition, these methods have a MAPE below 5%, which indicates the high efficiency of the methods.

Temperature dependencies observed in the temperate climate zone are also well amenable to processing by machine learning methods (Table 4). Immediately three methods (RF, LGBM and XGB) show high performance, while another 3 methods (DT, KNN, GB) also show high performance, although they are minimally inferior.

Forecasting the land surface temperature in the continental climate zone using machine learning methods demonstrates a fairly high accuracy (Table 5). All significant methods have an R^2 score above 94%. The gradient enhancement method has the best results. He is somewhat influenced by LGBT and XGB methods. The evaluation of MAPE and sMAPE metrics for continental and polar climatic zones was not carried out, since in these zones temperature values close to 0 are quite often observed, when extended to which, according to formulas 7 and 8, very high indicators are produced, which in fact do not reflect real situation.

For temperatures in the polar climate zone, the methods do not work very well (Table 6). The highest rates are observed among LGBM. R2 is 92%.

Since the time spent on creating a forecast plays a rather important role, it is advisable to choose faster methods, provided the same accuracy of forecasting. Table 7 shows the running time of each method for data from each climate zone. Comparing the results, we can conclude that the methods of decision trees and k-nearest neighbors work the fastest. Analyzing the forecasting accuracy, it can be concluded that the KNN, RF, GB, LGBM, XGB methods are quite effective for creating a forecast in the short and medium term. These methods are able to fairly accurately predict the average monthly temperature for the next decade. Tables 8 and 9 show the standard deviation, correlation coefficients, and centered root mean square error for each of the methods in each climate zone.

Table 3
ML methods' metrics evaluation for land surface temperature forecasting in zone B

Method	Metrics							
	MAE	MSE	RMSE	R2	MAPE, %	sMAPE, %	MBD	MedAE
NN	0.7624	0.8241	0.9078	0.9626	3.6509	3.5874	−0.017	0.7348
DT	0.4069	0.3311	0.5754	0.985	2.0901	2.072	−0.006	0.2769
RF	0.3826	0.3043	0.5516	0.9862	1.9779	1.9651	0.011	0.2779
KNN	0.4383	0.3461	0.5883	0.9843	2.1835	2.193	0.2083	0.3614
SVR	2.4464	7.337	2.7087	0.667	10.264	10.9008	2.3992	2.79
GB	0.4114	0.2941	0.5423	0.9867	2.0614	2.0606	0.1255	0.3287
AB	3.3319	12.209	3.4941	0.4459	14.539	15.7484	3.3247	3.5822
LGBM	0.3659	0.2488	0.4988	0.9887	1.8651	1.8592	−0.021	0.2597
XGB	0.3792	0.2954	0.5436	0.9866	1.9607	1.9435	−0.01	0.2425

Table 4
ML methods' metrics evaluation for land surface temperature forecasting in zone C

Method	Metrics							
	MAE	MSE	RMSE	R2	MAPE, %	sMAPE, %	MBD	MedAE
NN	0.6968	0.8324	0.9124	0.9431	4.9592	4.8864	0.056	0.5704

DT	0.4892	0.4466	0.6683	0.9695	3.6017	3.5947	0.04	0.3724
RF	0.4528	0.4193	0.6475	0.9713	3.347	3.3468	0.0782	0.3408
KNN	0.4826	0.4786	0.6918	0.9673	3.4636	3.5051	0.2661	0.3478
SVR	1.47	3.012	1.736	0.794	9.2742	9.7888	1.3526	1.4363
GB	0.4602	0.4422	0.665	0.9698	3.381	3.3721	0.1335	0.3555
AB	1.7784	4.4077	2.0995	0.6986	12.396	13.5777	1.741	1.7006
LGBM	0.4175	0.3879	0.6228	0.9735	3.0887	3.0719	0.0347	0.3008
XGB	0.4488	0.4137	0.6432	0.9717	3.3478	3.335	0.0420	0.3496

Table 5
ML methods' metrics evaluation for land surface temperature forecasting in zone D

Method	Metrics							
	MAE	MSE	RMSE	R2	MAPE, %	sMAPE, %	MBD	MedAE
NN	1.121	2.095	1.447	0.9725	−0.137	0.9426	1.121	2.095
DT	0.9041	1.391	1.179	0.9817	−0.16	0.7756	0.9041	1.391
RF	0.8834	1.286	1.134	0.9831	0.0497	0.7208	0.8834	1.286
KNN	0.944	1.355	1.164	0.9822	0.4538	0.8171	0.944	1.355
SVR	1.782	4.497	2.121	0.9409	1.1671	1.6315	1.782	4.497
GB	0.8221	1.126	1.061	0.9852	0.0787	0.6538	0.8221	1.126
AB	1.6938	4.3075	2.0754	0.9434	1.221	1.5527	1.6938	4.3075
LGBM	0.8315	1.1502	1.0724	0.9849	0.1413	0.608	0.8315	1.1502
XGB	0.822	1.2356	1.1116	0.9838	−0.154	0.6457	0.822	1.2356

Table 6
ML methods' metrics evaluation for land surface temperature forecasting in zone E

Method	Metrics							
	MAE	MSE	RMSE	R2	MAPE, %	sMAPE, %	MBD	MedAE
NN	1.121	2.095	1.447	0.9725	−0.137	0.9426	1.121	2.095
DT	0.9041	1.391	1.179	0.9817	−0.16	0.7756	0.9041	1.391
RF	0.8834	1.286	1.134	0.9831	0.0497	0.7208	0.8834	1.286
KNN	0.944	1.355	1.164	0.9822	0.4538	0.8171	0.944	1.355
SVR	1.782	4.497	2.121	0.9409	1.1671	1.6315	1.782	4.497
GB	0.8221	1.126	1.061	0.9852	0.0787	0.6538	0.8221	1.126
AB	1.6938	4.3075	2.0754	0.9434	1.221	1.5527	1.6938	4.3075
LGBM	0.8315	1.1502	1.0724	0.9849	0.1413	0.608	0.8315	1.1502
XGB	0.822	1.2356	1.1116	0.9838	−0.154	0.6457	0.822	1.2356

Table 7
Time spent on creating a forecast

Method	Time, sec				
	Zone A	Zone B	Zone C	Zone D	Zone E
NN	0.4	0.57	1.36	0.45	0.66
DT	0.01	0.01	0.01	0.01	0.01
RF	0.47	0.59	0.56	0.52	0.54
KNN	0.01	0.01	0.01	0.01	0.01
SVR	0.12	0.07	0.06	0.08	0.14
GB	0.21	0.24	0.21	0.19	0.26
AB	0.35	0.39	0.32	0.33	0.3
LGBM	0.07	0.10	0.08	0.07	0.07
XGB	0.09	0.12	0.14	0.13	0.12

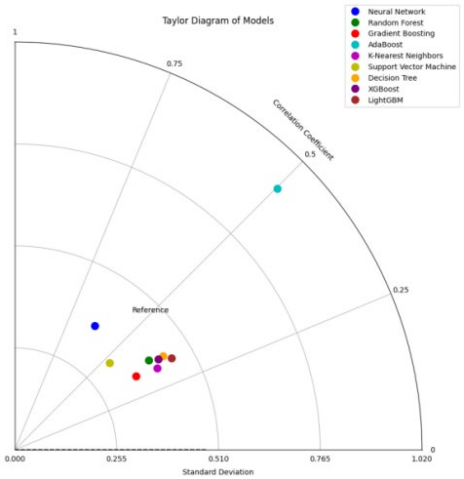
Table 8
Taylor statistics for ML methods in zones A, B and C

Method	Taylor statistics								
	Zone A			Zone B			Zone C		
	σ	r	E'	σ	r	E'	σ	r	E'
NN	0.369	0.542	0.446	3.975	0.992	0.554	3.207	0.982	0.672
DT	0.439	0.845	0.374	4.565	0.993	0.561	3.714	0.983	0.659
RF	0.403	0.832	0.313	4.589	0.993	0.541	3.698	0.986	0.635
KNN	0.411	0.868	0.244	4.609	0.993	0.582	3.679	0.986	0.676
SVR	0.323	0.737	0.718	3.637	0.986	2.494	3.013	0.977	1.534
GB	0.356	0.854	0.316	4.6	0.994	0.534	3.574	0.987	0.616
AB	0.928	0.709	1.192	3.839	0.988	3.388	3.929	0.955	2.097
LGBM	0.4546	0.863	0.246	4.703	0.994	0.499	3.668	0.987	0.603
XGB	0.426	0.846	0.365	4.576	0.993	0.531	3.685	0.986	0.628

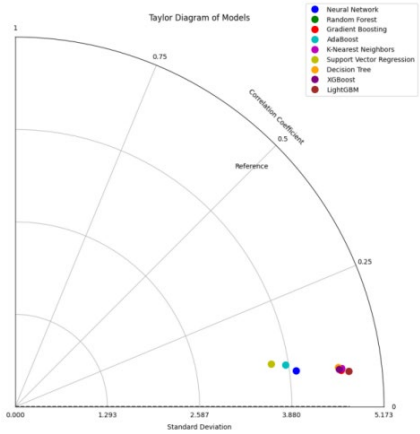
Table 9
Taylor statistics for ML methods in zones D and E

Method	Taylor statistics					
	Zone D			Zone E		
	σ	r	E'	σ	r	E'
NN	3.9749	0.9918	0.5540	2.1978	0.9324	1.2512
DT	4.5649	0.9927	0.5607	2.766	0.95	0.9195
RF	4.5891	0.9932	0.5415	2.7418	0.9536	0.8693
KNN	4.6088	0.9932	0.5821	2.6977	0.9574	0.8793
SVR	3.6370	0.9864	2.4939	2.6935	0.9494	2.3643
GB	4.6002	0.9938	0.5341	2.7258	0.9676	0.7720
AB	3.8389	0.9882	3.3879	1.9441	0.9437	2.2599
LGBM	4.7031	0.9944	0.4987	2.6913	0.9673	0.7147
XGB	4.5761	0.9934	0.5306	2.7396	0.9547	0.8717

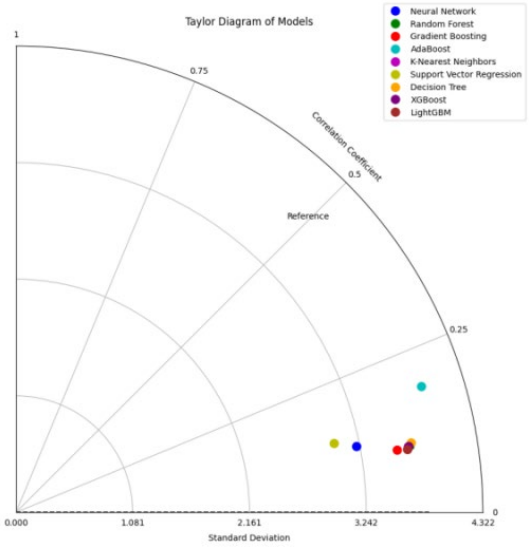
Figure 1 shows Taylor diagrams for the considered machine learning methods for each of the climate zones.



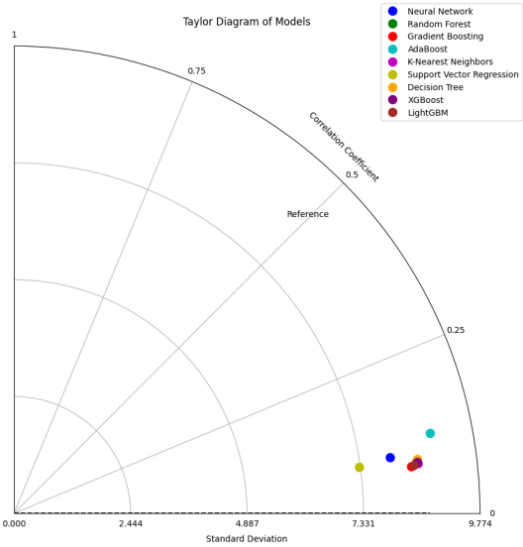
a) Zone A



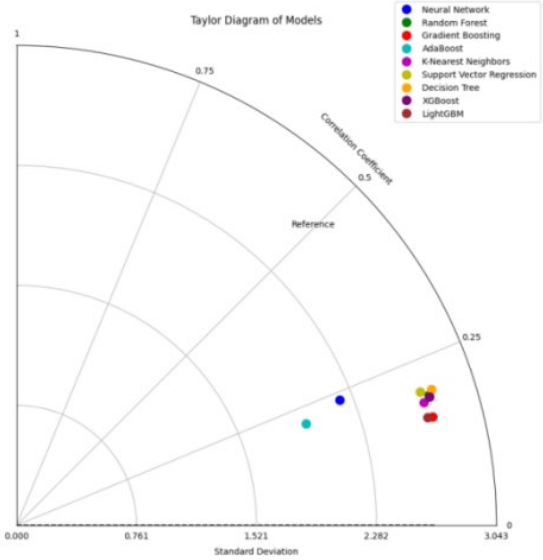
b) Zone B



c) Zone C



d) Zone D



e) Zone E

Figure 1: Taylor diagrams for the considered machine learning methods.

5. Conclusions

The conducted research made it possible to identify the most effective methods for forecasting the temperature of the Earth's surface in terms of the accuracy of the forecast and the time spent in each of the climatic zones. Analysis of the forecast, taking into account climatic zoning, allows to more clearly determine the patterns of individual territories and make a more accurate forecast. The proposed approach makes it possible to monitor the main trends of climatic changes in the context of changes in the temperature of the Earth's surface in the short- and medium-term perspectives. The proposed machine learning methods are able to make an accurate and quick forecast of the main trends in the change of the average monthly temperature of the Earth's surface for the next decade. Evaluation of machine learning methods was carried out on the basis of metrics. The values of standard deviation, correlation coefficient and centered mean squared error for each of the methods were also calculated. To visualize the effectiveness of the methods, Taylor diagrams were constructed. This research makes it possible to form a basis for further study of changes in climatic indicators in the context of individual territories and the search for the most appropriate machine learning methods for forecasting climatic changes, taking into account climatic zoning.

Acknowledments

This work was supported by the project "Earth Observation for Early Warning of Land Degradation at European Frontier (EWALD)" under the European Union's Framework Programme for Research and Innovation Horizon Europe – the Framework Programme for Research and Innovation (2021-2027), Grant Agreement No. ID 101086250.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check; DeepL Translate in order to: some phrases translation into English. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Hryhoruk, S. Grygoruk, N. Khrushch, T. Hovorushchenko. Using non-metric multidimensional scaling for assessment of regions' economy in the context of their sustainable development. CEUR-WS. (2020). Vol. 2713. 315-333.

[2] S. Yıldırım, S.H. Bostancı, D.Ç. Yıldırım. Parameters for the Study of Climate Refugees. In: P. Singh, B. Ao, A. Yadav (eds) Global Climate Change and Environmental Refugees. Springer, Cham. (2023). doi:10.1007/978-3-031-24833-7_11.

[3] C. O. de Burgh-Day, T. Leeuwenburg. Machine learning for numerical weather and climate modelling: a review, Geosci. Model Dev., 16, 6433–6477, (2023).

[4] L. Chen, B. Han, X. Wang, J. Zhao, W. Yang, Z. Yang. Machine Learning Methods in Weather and Climate Applications: A Survey. Appl. Sci. (2023), 13, 12019. doi:10.3390/app132112019.

[5] T. Hovorushchenko, V. Alekseiko. Land surface temperature forecasting in the context of the development of sustainable cities and communities. Computer Systems and Information Technologies, 3, (2024). 6–12. doi:10.31891/csit-2024-3-1.

[6] D. Fister, J. Pérez-Aracil, C. Peláez-Rodríguez, J. Del Ser, S. Salcedo-Sanz. Accurate long-term air temperature prediction with Machine Learning models and data reduction techniques, Applied Soft Computing, Volume 136, (2023). 110118, ISSN 1568-4946.

[7] M. Jamei, M. Karbasi, M. Ali, A. Malik, X. Chu, Z. M. Yaseen, A novel global solar exposure forecasting model based on air temperature: Designing a new multi-processing ensemble deep

learning paradigm. Expert Systems with Applications, Volume 222, (2023), 119811, ISSN 0957-4174, doi:10.1016/j.eswa.2023.119811.

[8] C. B. Pande, J. C. Egbueri, R. Costache, L. M. Sidek, Q. Wang, F. Alshehri, N. Md Din, V. K. Gautam, S. C. Pal. Predictive modeling of land surface temperature (LST) based on Landsat-8 satellite data and machine learning models for sustainable development, Journal of Cleaner Production, Volume 444, (2024). 141035, ISSN 0959-6526.

[9] F. Di Nunno, S. Zhu, M. Ptak, M. Sojka, F. Granata. A stacked machine learning model for multi-step ahead prediction of lake surface water temperature, Science of The Total Environment, Volume 890, (2023). 164323, ISSN 0048-9697. doi:10.1016/j.scitotenv.2023.164323.

[10] O. E. Adeyeri, A. H. Folorunsho, K. I. Ayegbusi, V. Bobde, T. E. Adeliyi, C. E. Ndehedehe, A. A. Akinsanola. Land surface dynamics and meteorological forcings modulate land surface temperature characteristics, Sustainable Cities and Society, Volume 101, (2024), 105072, ISSN 2210-6707, doi:10.1016/j.scs.2023.105072.

[11] N. Gupta, B. H. Aithal. Urban land surface temperature forecasting: a data-driven approach using regression and neural network models. Geocarto International, 39(1). (2024). https://doi.org/10.1080/10106049.2023.2299145.

[12] L. Tian, Y. Tao, M. Li, C. Qian, T. Li, Y. Wu, F. Ren . Prediction of Land Surface Temperature Considering Future Land Use Change Effects under Climate Change Scenarios in Nanjing City, China. Remote Sensing. 15(11):2914. (2023). doi:10.3390/rs15112914.

[13] Kaggle. Globallandtemperature. (2018). https://www.kaggle.com/datasets/sambapython/globallandtemperature.

[14] List of countries by climate zone and average yearly temperatures. (2024) https://weatherandclimate.com/countries#google_vignette

[15] O. Pavlova, V. Alekseiko. The concept of an information system for forecasting the temperature regime of the earth's surface based on machine learning. Computer Systems and Information Technologies, №2, (2024). pp. 6–13. doi:10.31891/csit-2024-2-1

[16] S. Sharafi, M. Mohammadi Ghaleni. Revealing accuracy in climate dynamics: enhancing evapotranspiration estimation using advanced quantile regression and machine learning models. Appl Water Sci 14, 162. (2024). doi:10.1007/s13201-024-02211-5

[17] A. Nailman. Comparing machine learning algorithms for regression. Machine Learning Models. (2024, May 31). https://machinelearningmodels.org/comparing-machine-learning-algorithms-for-regression/

[18] B. Lefoula, A. Hebal, , & D. Bengora. Performance of machine learning methods for modeling reservoir management based on irregular daily data sets: a case study of Zit Emba dam. Earth Science Informatics, 17(1), (2023) pp. 145–161. doi:10.1007/s12145-023-01160-y.

[19] A. Nailman. Supervised Machine Learning types: Exploring the different approaches. Machine Learning Models. (2024, May 28). https://machinelearningmodels.org/supervised-machine-learning-types-exploring-the-different-approaches/.

[20] J. Chen. Analysis of Statistic Metrics in Different Types of Machine Learning. Highlights in Science, Engineering and Technology, 88, pp. 182–188. (2024). doi:10.54097/c4mz2q66.

[21] V. Plevris, G. Solorzano, N. Bakas, M. Ben Seghier. Investigation of performance metrics in regression analysis and machine learning-based prediction models. The 8th European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS Congress 2022. 5 – 9 June 2022, Oslo, Norway. (2022). doi:10.23967/eccomas.2022.155.

[22] B. Wohlwend. Regression model evaluation metrics: R-Squared, Adjusted R-Squared, MSE, RMSE, and MAE. Medium. (2023, August 12). https://medium.com/@brandon93.w/regression-model-evaluation-metrics-r-squared-adjusted-r-squared-mse-rmse-and-mae-24dcc0e4cbd3