The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 1, February 2025
https://doi.org/10.33022/ijcs.v14i1.4593

Optimizing Climate Forecasts Across 16 Zones Using Regression-Based Machine Learning Models

Ardin[1], Indra Budi[2]
ardin21@ui.ac.id[1], indra@cs.ui.ac.id[2]
[1,2] Faculty of Computer Science, University of Indonesia

Article Information

Received : 29 Dec 2024
Revised : 6 Jan 2025
Accepted : 29 Jan 2025

Keywords

Decadal Rainfall, Seasonal Zone, Machine Learning

Abstract

The XYZ Climatology Station faces challenges in improving the accuracy of decadal rainfall forecasts, with an average achievement of 57.4% in 2022 and 58.8% in 2023, below the organizational performance target of 70% accuracy as set in its strategic objectives. This study aims to develop machine learning-based predictive models for 16 climate zones to enhance forecast accuracy. Five regression algorithms—Multiple Linear Regression, Support Vector Regression, Extra Trees Regression, Random Forest Regression, and Decision Tree Regression—were tested under two scenarios: input variable variations (VR) and time series data length (TS). Results showed that the VR scenario increased average accuracy to 71.7% (2022) and 69.4% (2023), while the TS scenario achieved 73.1% (2022) and 72.6% (2023). Support Vector Regression and Extra Trees Regression demonstrated the best performance in most zones. These models are expected to be operationalized to improve climatological information services and better meet public and stakeholder needs.

A. Introduction

Indonesia, located along the equator and surrounded by the Indian and Pacific Oceans, experiences a complex climate shaped by global phenomena such as the El Niño Southern Oscillation (ENSO), the Indian Ocean Dipole (IOD), the Asia-Australia monsoon circulation, the Intertropical Convergence Zone (ITCZ), and its diverse topography [1]. To address these challenges, the Indonesian government established an institution tasked with delivering high-quality climatological services to support public safety and welfare [2]. Among these institutions, the XYZ Climatology Station set a target forecast accuracy of 70% for rainfall predictions in 2022 and 2023 [3]. However, the station faced significant challenges due to climate variability, exacerbated by anomalous climate events such as El Niño and La Niña during these years [4], [5]. To mitigate these challenges, XYZ Climatology Station employs multiple climate models and integrates their outputs. This approach aims to reduce the uncertainties and errors typically associated with single-model limitations and the complexity of climate variability [6], [7]. By combining forecasts from various models, each with its own strengths and weaknesses, this ensemble method seeks to improve the accuracy of climate predictions [6], [8], [9]. Traditional ensemble techniques, such as the Average Ensemble or Weighted Average Ensemble, have been used, but these methods have yielded suboptimal results at XYZ Climatology Station in practice. An evaluation of rainfall forecasts, comparing predicted values with actual observational data using the Proportion of Correct (PC) metric, revealed that in 2022 and 2023, many zones did not meet the accuracy target. The average accuracy across all zones was 57.4% in 2022 and 58.8% in 2023, significantly below the target set.

Research on ensemble methods using outputs from different climate models to produce more accurate climate forecasts has been widely conducted, yet often using different ensemble techniques, such as regression via machine learning algorithms. Machine learning is effective for addressing complex problems that traditional methods struggle to solve, as it can uncover relevant insights [10], [11]. The use of appropriate algorithms for decision-making has been increasingly developed, enhancing autonomy and control [12]. Additionally, these algorithms improve the accuracy of rainfall data analysis, resulting in more accurate forecasts [13]. Explorations of machine learning with 21 datasets from global models (NEX-GDDP) and 13 CMIP6 models using algorithms such as Multiple Linear Regression (MLR), Support Vector Machine (SVM), Extra Tree Regressor (ETR), Random Forest (RF), and Long Short-Term Memory (LSTM) have successfully reduced prediction uncertainties and enhanced the accuracy of climate models at local scales. Li et al. [14] compared ensemble strategies using the average arithmetic mean (AM) and linear regression among ensemble members, with RF modeling. Their study found that RF in multi-model ensemble processing produced more accurate climate projections, improving the precision of rainfall and temperature predictions and identifying spatial differences in greater detail. Other studies proposed multi-model ensemble techniques to improve climate projection accuracy by combining 36 outputs from General Circulation Models (GCMs) using algorithms such as Artificial Neural Networks (ANN), K-Nearest Neighbour (KNN), SVM, and Relevance Vector Machine (RVM), which significantly enhanced climate projection accuracy. In general, machine learning algorithms can correct errors and improve the accuracy of rainfall data analysis, particularly through optimized regression models, thus yielding more accurate forecasts [13].

Building on previous studies and addressing the challenges in climatological information services, particularly at Climatology Stations, this research aims to explore several machine learning-based regression algorithms through various experimental scenarios applied to climate model forecast datasets. The objective is to develop an optimal predictive model that enhances rainfall forecast accuracy across 16 climate zones at XYZ Climatology Station. This study is expected not only to help achieve organizational targets by improving forecast accuracy but also to provide more reliable decadal rainfall data for each zone, thereby offering significant value to the public and stakeholders who depend on this information.

B. Research Method

This research focuses on the exploration of various machine learning-based regression algorithms, combined with different experimental scenarios applied to climate model forecast datasets. The aim is to obtain an optimal configuration that produces the best predictive model, capable of improving rainfall forecast accuracy across the 16 zones under the responsibility of XYZ Climatology Station. In developing this predictive model, several research instruments are required, including:

1. Data

This study utilizes two categories of climate parameter data: rainfall observation data, which is historical rainfall data obtained from direct measurements, and data from eight climate models, consisting of historical outputs from multiple model forecasts developed and used in operational settings, covering the period from 1991 to 2023. These two datasets will be used for training the regression models, with each serving as the target and input variables, as shown in Table 1.

Table 1. Main Variable Data

No	Category	Type	Description
1	Rainfall Observation Data	Numeric	Daily precipitation data from direct measurements
2	Climate Model Data (Multiple Model Forecast):		
	1) rawECMWF	Numeric	Multi-model forecast outputs of decadal rainfall (ten-day accumulated rainfall)
	2) corECMWF	Numeric	
	3) rawCFS	Numeric	
	4) corCFS	Numeric	
	5) WAR	Numeric	
	6) ARM	Numeric	
	7) InaMMEv1	Numeric	
	8) InaMMEv1	Numeric	

2. Research Tools

The research utilizes a range of software and libraries for data processing and model development. Microsoft Excel is employed for converting daily rainfall observations into decadal (ten-day) periods and constructing the dataset, taking advantage of its data processing. The Pandas library in Python is used for efficient data manipulation and analysis within DataFrame structures. For regression modeling, the Scikit-Learn library, a popular Python tool for machine learning, is applied. All programming tasks are carried out using Google Colab, a cloud-based platform that allows for direct Python code writing and execution in the browser.

The research experiment process involves several stages: data collection, preprocessing, modeling, evaluation, and validation. These stages are carried out in sequence, following the workflow illustrated in Figure 1.

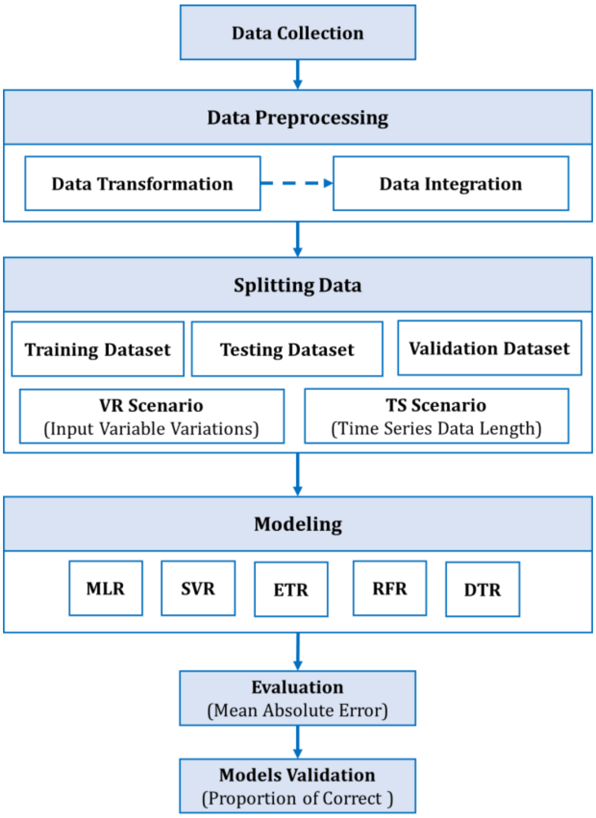


Figure 1. Research Workflow

Each step and process in this research, as illustrated in Figure 1, will be explained in more detail as follows:

1. **Data Collection**
   The research relies on two primary types of data: rainfall observation data and outputs from various climate models. These datasets are sourced directly from the XYZ Climatology Station, which is responsible for providing climate forecasts and conducting climate observations across 16 distinct climate zones. This data represents historical observations and operational forecasting services provided by the station.
2. **Data Preprocessing**

The data preprocessing steps in this research are relatively simple, as the data has already been well-documented through operational activities. The preprocessing tasks include:

a) Data Transformation, which involves converting daily rainfall data into decadal (10-day) rainfall data by aggregating the data for each zone to meet the research needs.
b) Data Integration, where the transformed data from various sources is combined into a unified dataset in DataFrame format. This consolidated dataset is then prepared for modeling and further analysis.

3. **Dataset Partitioning and Experimental Scenarios**

The dataset used in this study is divided into three parts:

a) Training Dataset: This dataset is used to train the model, enabling it to recognize patterns, understand relationships between variables, and extract knowledge from the data. It covers a maximum period of 30 years, from 1991 to 2020. The study also incorporates two experimental scenario designs that influence the training dataset used in modeling:
   - First scenario: Variation in input variables (VR), consisting of three experiments—VR-1, VR-2, and VR-3—focusing on the addition of input variables.
   - Second scenario: Variation in time series data length (TS), involving six experiments—TS-1 to TS-6—emphasizing differences in historical data length.

   These approaches aim to explore various regression modeling configurations to identify the most optimal and reliable predictive model. Details of both scenarios are presented in Table 2.

Table 2. Experimental Scenarios for Model Training

Scenario Name	Input Variables	Series Data
VR-1	rawECMWF, corECMWF, rawCFS, corCFS	All series data
VR-2	rawECMWF, corECMWF, rawCFS, corCFS, WAR, ARM	All series data
VR-3	rawECMWF, corECMWF, rawCFS, corCFS, WAR, ARM, InaMMEv1, InaMMEv2	All series data
TS-1	All variabel input data	5 Years (2016- 2020)
TS-2	All variabel input data	10 Years (2011-2020)
TS-3	All variabel input data	15 Years (2006-2020)
TS-4	All variabel input data	20 Years (2001-2020)
TS-5	All variabel input data	25 Years (1996-2020)
TS-6	All variabel input data	30 Years (1991-2020)

b) Testing Dataset: Used to assess the performance and effectiveness of the trained model by evaluating its ability to generalize patterns to unseen data, thereby identifying the most optimal model. This dataset consists of data for a single year, 2021.
c) Validation Dataset: Utilized to evaluate the top-performing model from the testing phase, ensuring its reliability in addressing real-world challenges relevant to the research. This dataset spans two years, 2022 and 2023, which are the focal periods of this study.

4. **Modeling**

   The algorithms employed in this research experiment include:

   a) Multiple Linear Regression (MLR) is a widely used regression analysis method to explain the relationship between one dependent variable and multiple independent variables through a linear equation [15]. This method is commonly applied in climate studies for downscaling and impact analysis [16], [17].
   b) Support Vector Regression (SVR) is a non-linear regression algorithm that maps low-dimensional data into a high-dimensional feature space using kernel functions. This technique is frequently utilized in climate change and hydrological analyses [18], [19].
   c) Extra Trees Regression (ETR) is a variation of decision tree algorithms that incorporates randomness in its formation [20]. Unlike Random Forest (RF), ETR: (1) does not use bootstrapping and trains each tree on the entire training dataset and (2) selects split points randomly instead of determining the optimal split. The split with the highest score among the random options is chosen. This approach generates unique decision trees for each sample, helping to mitigate overfitting [21], [22].
   d) Random Forest Regressor (RFR) utilizes an ensemble of decision trees to prevent overfitting and handle various types of input variables. This algorithm produces independent trees and makes predictions based on non-parametric statistical regression with randomization elements [22]. The final prediction is derived by averaging the outputs of all trees.
   e) Decision Tree Regression (DTR) is a predictive method for regression tasks that splits the dataset into subsets based on feature values, constructing a decision tree with branches representing decisions and leaves representing predictions. DTR excels at handling non-linear and multivariate data without requiring assumptions about the data distribution [23], [24].

5. **Evaluation**

   Model evaluation involves assessing the performance of machine learning models using metrics such as Mean Absolute Error (MAE) to ensure accurate and reliable predictions. MAE is widely applied in research, including in the field of climatology [25], [26].

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|$$

6. **Validation**

   This study employs the Proportion of Correct (PC) metric, a standard operational metric used by the XYZ Climatology Station for evaluating and verifying climate forecasts. The use of this metric ensures that the model's performance results are interpretable within the organization's operational context. PC is a simple and intuitive metric that measures the accuracy of categorical forecasts. As noted by Jolliffe [27] and cited in Muharsyah [28], PC is a commonly used verification technique for evaluating forecast accuracy in climatology.

ij		Observation				Summation
		1	2	…j…	k	$\sum P_{ij}$
Forecast	1	P_{11}	P_{12}	P_{1j}	P_{1k}	$\sum P_{ij}$
	2	P_{21}	P_{22}	P_{2j}	P_{2k}	$\sum P_{ij}$
	…i…	P_{i1}	P_{i2}	P_{ij}	P_{ik}	$\sum P_{ij}$
	k	P_{k1}	P_{k2}	P_{kj}	P_{kk}	$\sum P_{ij}$
Summation		$\sum P_{ij}$	$\sum P_{ij}$	$\sum P_{ij}$	$\sum P_{ij}$	1

Figure 2. Contingency Table of Matching Pairs Between Forecast and Observation

C. Result and Discussion

Data Collection

The climate parameters used in this study are divided into two categories: observational rainfall data and data from eight climate models.

- Observational Rainfall: This data is collected through direct measurements by the staff of the XYZ Climatology Station, which records daily rainfall at specific locations representing each study zone. The daily data is then processed into decadal time scales (Figure 3).

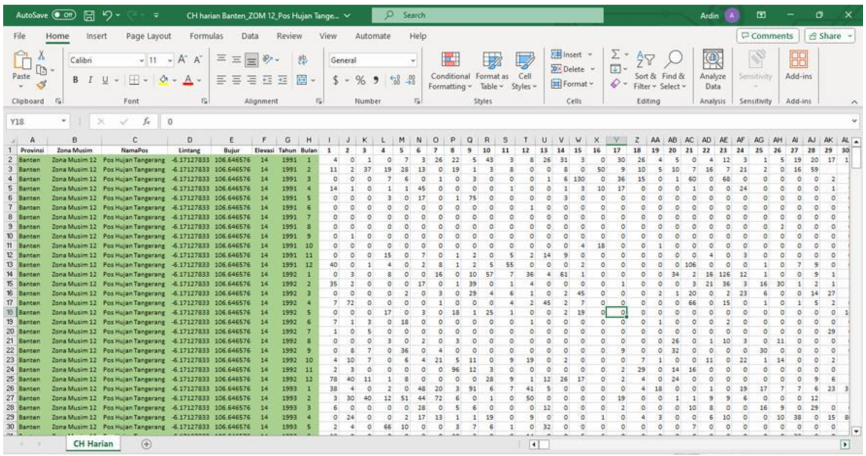
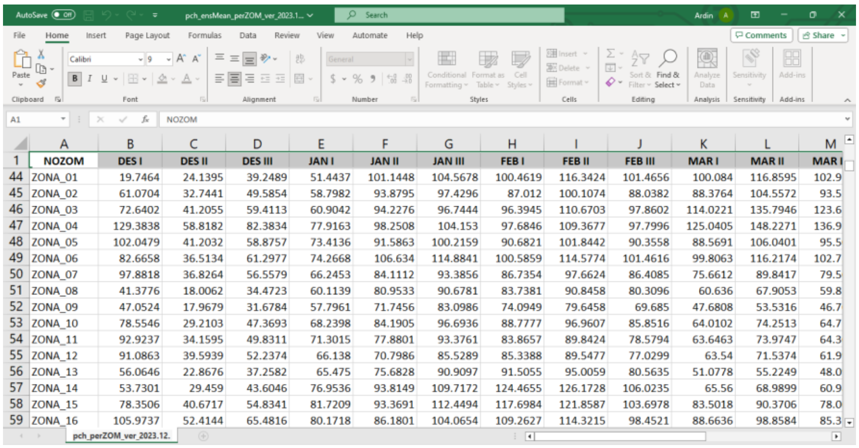


Figure 3. Screenshot of the Observational Rainfall Dataset

- Climate Model Data: This data is derived from outputs of eight climate models that provide rainfall forecasts on a decadal time scale. These data are historical operational results from the XYZ Climatology Station and serve as representations of climate predictions (Figure 4).



	NOZOM	DES I	DES II	DES III	JAN I	JAN II	JAN III	FEB I	FEB II	FEB III	MAR I	MAR II	MAR I
44	ZONA_01	19.7464	24.1395	39.2489	51.4437	101.1448	104.5678	100.4619	116.3424	101.4656	100.084	116.8595	102.9
45	ZONA_02	61.0704	32.7441	49.5854	58.7982	93.8795	97.4296	87.012	100.1074	88.0382	88.3764	104.5572	93.5
46	ZONA_03	72.6402	41.2055	59.4113	60.9042	94.2276	96.7444	96.3945	110.6703	97.8602	114.0221	135.7946	123.6
47	ZONA_04	129.3838	58.8182	82.3834	77.9163	98.2508	104.153	97.6846	109.3677	97.7996	125.0405	148.2271	136.9
48	ZONA_05	102.0479	41.2032	58.8757	73.4136	91.5863	100.2159	90.6821	101.8442	90.3558	88.5691	106.0401	95.5
49	ZONA_06	82.6658	36.5134	61.2977	74.2668	106.634	114.8841	100.5859	114.5774	101.4616	99.8063	116.2174	102.7
50	ZONA_07	97.8818	36.8264	56.5579	66.2453	84.1112	93.3856	86.7354	97.6624	86.4085	75.6612	89.8417	79.5
51	ZONA_08	41.3776	18.0062	34.4723	60.1139	80.9533	90.6781	83.7381	90.8458	80.3096	60.636	67.9053	59.8
52	ZONA_09	47.0524	17.9679	31.6784	57.7961	71.7456	83.0986	74.0949	79.6458	69.685	47.6808	53.5316	46.7
53	ZONA_10	78.5546	29.2103	47.3693	68.2398	84.1905	96.6936	88.7777	96.9607	85.8516	64.0102	74.2513	64.7
54	ZONA_11	92.9237	34.1595	49.8311	71.3015	77.8801	93.3761	83.8657	89.8424	78.5794	63.6463	73.9747	64.3
55	ZONA_12	91.0863	39.5939	52.2374	66.138	70.7986	85.5289	85.3388	89.5477	77.0299	63.54	71.5374	61.9
56	ZONA_13	56.0646	22.8676	37.2582	65.475	75.6828	90.9097	91.5055	95.0059	80.5635	51.0778	55.2249	48.0
57	ZONA_14	53.7301	29.459	43.6046	76.9536	93.8149	109.7172	124.4655	126.1728	106.0235	65.56	68.9899	60.9
58	ZONA_15	78.3506	40.6717	54.8341	81.7209	93.3691	112.4494	117.6984	121.8587	103.6978	83.5018	90.3706	78.0
59	ZONA_16	105.9737	52.4144	65.4816	80.1718	86.1801	104.0654	109.2627	114.3215	98.4521	88.6636	98.8584	85.3

Figure 4. Screenshot of the Climate Model Output Rainfall Dataset

Data Preprocessing

The data preprocessing in this study consists of two main steps: transformation and integration. The transformation step involves converting the daily rainfall data into decadal rainfall data, which is the accumulation of daily rainfall over a ten-day period, resulting in 36 decadal periods in a year [1]. Data integration combines separate files into a complete dataset in DataFrame format, ensuring that the data is ready for regression modeling. This DataFrame structure facilitates data management, analysis, and model training using the Pandas library in Python, as shown in Figure 5.

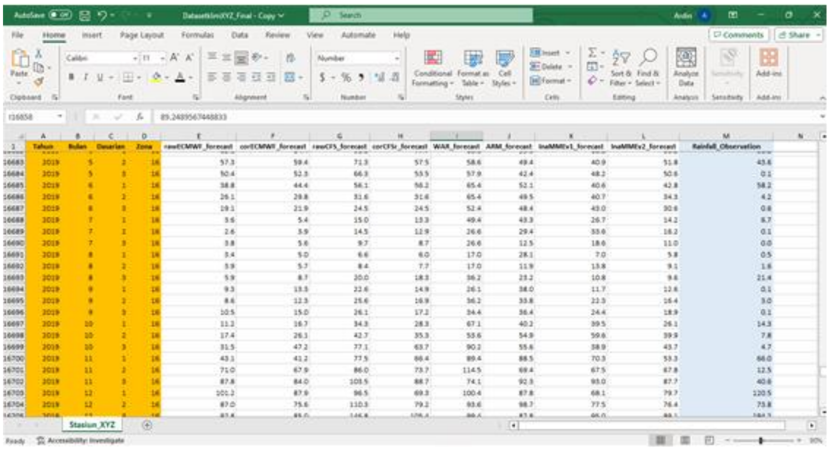


Figure 5. Screenshot of the Transformed and Integrated Dataset Structure

Dataset Partitioning

The Training Dataset consists of decadal rainfall data for the period 1991–2020, encompassing 1,080 decadal data points for each input and target variable. This dataset is used to train the models based on two experimental scenarios, as described in Table 2. Meanwhile, the Testing Dataset contains decadal rainfall data for one year (36 data points), used to evaluate model performance. The Validation Dataset spans two years (72 data points) and is used to assess the model's ability to address the challenges posed by the research period.

Modeling and Evaluation

The regression modeling experiments are conducted based on two main scenarios: variations in input variables (VR) and the time series length (TS). Five regression algorithms are used for modeling: MLR, SVR, DTR, ETR, and RFR. The trained models are then tested across 16 zones using the MAE metric to measure prediction errors. Each model is evaluated under various scenarios to determine the best predictive performance for each zone. The results are displayed in Figures 6 and 7.

Table 3. Evaluation of Predictive Model Performance by Zone for VR Scenario

Models		Zona															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
VR-1	MLR	35.3	41.1	43.8	52.6	33.6	36.8	38.5	28.7	30.3	30.4	33.2	35.2	30.6	30.9	38.2	41.7
	SVR	35.7	38.8	41.5	53.5	34.3	34.6	38.8	28.3	31.3	30.9	34.2	36.4	30.1	31.8	41.0	41.7
	DTR	68.5	58.2	77.2	85.7	56.9	65.4	45.7	39.2	41.9	53.9	49.8	53.9	34.0	39.3	47.3	60.8
	ETR	39.4	46.1	52.6	54.1	36.0	38.7	36.7	31.0	27.6	29.4	32.3	32.7	30.3	30.1	35.3	42.4
	RFR	39.7	46.1	51.5	52.2	35.5	40.3	39.7	31.2	32.1	31.5	33.9	35.3	29.8	31.2	35.5	43.6
VR-2	MLR	35.3	41.1	44.2	53.1	33.3	36.6	38.4	27.8	29.6	29.7	33.1	35.8	30.0	30.2	38.3	42.0
	SVR	35.5	38.9	41.7	53.3	34.1	34.0	38.8	28.4	31.6	30.9	34.1	36.7	29.4	31.0	40.8	40.5
	DTR	60.8	58.6	69.7	89.0	50.0	64.2	67.4	45.0	39.9	34.9	42.7	46.6	36.5	51.6	51.0	45.7
	ETR	41.3	45.9	48.4	60.3	37.3	38.5	37.8	30.3	29.8	29.9	31.0	33.0	28.8	31.6	35.6	42.5
	RFR	41.4	46.0	45.5	57.4	37.3	39.9	41.7	31.6	32.3	30.9	32.2	36.3	29.8	31.6	38.6	44.4
VR-3	MLR	35.0	41.3	43.5	52.1	35.5	37.8	38.9	32.6	30.0	31.1	35.1	38.9	31.8	32.8	36.0	44.4
	SVR	35.9	39.5	41.8	53.9	35.1	35.0	39.5	28.8	29.6	30.9	34.0	36.5	28.6	32.2	37.9	43.4
	DTR	49.6	55.7	66.1	103.3	58.1	72.3	57.5	40.7	33.0	36.8	44.6	40.7	38.8	43.3	49.4	57.3
	ETR	43.7	47.9	49.0	62.4	38.6	42.0	40.6	32.9	32.3	31.0	32.2	35.8	30.3	31.3	36.4	41.6
	RFR	43.7	47.9	43.3	64.6	40.9	47.5	44.1	35.2	31.7	33.7	33.2	37.5	31.0	30.6	37.9	43.2

The evaluation results in Figure 6 show that the ETR algorithm consistently outperforms others in predicting decadal rainfall with low prediction errors. This algorithm excels in seven zones, namely Zones 7, 9, 10, 12, 14, and 15 under the VR-1 scenario, as well as Zone 11 under the VR-2 scenario. ETR's dominance highlights its ability to reliably capture rainfall patterns across different zones. Meanwhile, SVR algorithm performs strongly, achieving the best results in five zones under different scenarios. SVR outperforms in Zones 2 and 3 under the VR-1 scenario, Zones 6 and 16 under VR-2, and Zone 13 under VR-3. The flexibility of SVR across various scenarios demonstrates its potential to handle the complexity of rainfall variability.

MLR model also proves to be reliable, with its best performance in four zones: Zones 5 and 8 under VR-2, and Zones 1 and 4 under VR-4. MLR’s success in these zones underscores its ability to perform well in specific situations, although its overall scope is more limited compared to ETR and SVR. In contrast, DTR and RFR algorithms did not show superior performance in any of the tested zones. This indicates that these two algorithms are less competitive compared to ETR, SVR, and MLR in the context of decadal rainfall modeling.

Table 4. Evaluation of Predictive Model Performance by Zone for TS Scenario

Models		Zone 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
TS-1	MLR	37.2	45.8	47.4	52.0	38.6	40.2	43.1	43.6	42.8	36.7	42.5	43.6	37.8	39.8	41.8	47.9
	SVR	37.8	39.9	56.1	55.7	35.1	35.9	40.2	30.5	30.2	31.6	33.6	35.4	31.8	32.5	37.0	41.0
	DTR	53.3	85.2	76.0	86.1	76.9	82.1	49.4	61.6	58.5	53.7	46.7	50.7	63.4	48.4	59.2	58.4
	ETR	40.9	48.1	54.9	62.0	42.4	42.5	41.8	39.3	40.1	33.7	32.2	37.2	41.2	31.3	34.8	41.7
	RFR	43.7	49.9	43.6	63.0	46.8	45.1	42.0	42.9	42.1	36.6	33.2	37.0	41.6	30.6	38.8	43.3
TS-2	MLR	35.6	43.7	45.7	52.4	37.7	38.2	41.3	38.5	36.8	33.7	39.5	41.7	34.6	36.5	39.5	46.4
	SVR	36.9	39.8	51.7	54.6	35.4	35.3	39.7	30.1	29.9	31.3	33.5	35.7	31.7	32.5	36.9	40.9
	DTR	63.0	61.5	73.0	85.1	60.3	53.0	49.7	36.8	39.0	56.2	43.7	50.1	47.1	41.0	53.5	66.0
	ETR	43.0	49.6	54.2	67.3	40.9	43.8	41.6	34.8	33.2	32.9	31.0	37.1	34.1	32.2	34.9	42.1
	RFR	41.4	50.1	42.8	63.9	41.4	44.1	42.1	37.8	35.4	35.3	33.7	38.2	33.1	31.6	36.4	42.3
TS-3	MLR	35.3	42.2	44.4	52.9	36.9	38.1	40.0	34.6	32.0	32.4	36.6	40.1	31.5	33.9	37.3	44.6
	SVR	36.3	39.5	47.5	53.8	35.4	35.1	39.3	29.3	30.1	30.8	33.7	36.1	31.8	32.3	37.4	41.6
	DTR	63.5	73.1	65.4	97.1	51.6	66.7	72.8	46.3	44.1	44.4	40.1	41.0	35.6	58.3	57.9	54.7
	ETR	41.8	47.6	49.9	65.4	40.1	42.5	41.9	33.0	32.8	32.2	32.5	37.9	29.8	32.9	36.4	44.5
	RFR	41.4	47.4	46.7	62.8	42.8	47.8	48.9	33.7	33.5	34.3	34.6	36.0	32.3	34.9	37.9	45.0
TS-4	MLR	35.4	42.0	44.0	52.6	36.7	38.5	40.0	34.8	32.2	32.7	36.8	40.2	31.4	34.2	37.3	44.7
	SVR	36.6	39.6	41.9	53.8	35.4	35.1	39.5	29.3	29.8	31.2	33.7	36.3	31.9	32.5	37.7	41.7
	DTR	59.6	80.6	51.1	78.9	61.5	90.7	83.5	51.7	47.5	45.0	51.6	61.1	40.6	50.4	62.6	60.7
	ETR	47.7	47.0	51.1	66.5	41.4	39.4	43.7	33.9	32.7	31.0	32.8	34.5	29.3	31.6	35.7	43.0
	RFR	45.8	51.5	42.0	65.6	41.4	48.2	46.8	34.0	34.3	33.7	35.1	37.2	31.2	34.8	38.6	46.3
TS-5	MLR	34.9	41.0	43.2	51.9	35.2	37.5	38.6	32.8	30.4	31.1	35.1	38.9	30.6	33.1	36.2	43.2
	SVR	36.9	39.4	47.1	54.0	35.4	35.0	39.5	29.1	29.7	31.1	34.0	36.3	32.1	32.4	38.4	42.0
	DTR	56.0	75.2	57.5	101.8	61.7	76.4	79.8	45.4	43.1	36.9	61.8	52.1	49.4	46.5	63.0	54.2
	ETR	45.8	47.9	46.1	62.6	43.0	41.0	41.5	32.0	31.7	30.3	33.2	36.5	30.9	33.2	35.8	44.2
	RFR	42.9	48.3	41.9	63.7	40.2	46.4	45.8	32.9	30.7	35.3	35.7	38.0	30.7	34.8	38.5	48.0
TS-6	MLR	35.0	41.3	43.5	52.1	35.5	37.8	38.9	32.6	30.0	31.1	35.1	38.9	30.3	32.8	36.0	43.2
	SVR	35.9	39.5	43.3	53.9	35.5	35.1	39.5	28.8	29.6	30.9	34.0	36.5	31.8	32.2	37.9	41.6
	DTR	49.6	55.7	66.1	103.3	58.1	72.3	57.5	40.7	33.0	36.8	44.6	40.7	38.8	43.3	49.4	57.3
	ETR	43.7	47.9	49.0	62.4	38.6	42.0	40.6	32.9	32.3	30.6	34.0	35.8	30.1	35.7	37.5	43.4
	RFR	43.7	47.9	41.8	64.6	40.9	47.5	44.1	35.2	31.7	33.4	35.2	37.5	31.0	37.7	39.8	44.4

The evaluation results in Figure 7 show that the ETR algorithm exhibits superior performance in predicting decadal rainfall with low error rates across six different zones. ETR achieved the best performance in Zone 15 under the TS-1 scenario, Zone 11 under TS-2, Zones 12 and 13 under TS-4, and Zone 10 under TS-5 and TS-6. This excellence indicates that ETR consistently captures rainfall patterns across various data scenarios. Meanwhile, SVR algorithm also performs competitively, achieving the best results in six other zones. SVR successfully predicts with high accuracy in Zone 5 under TS-1, Zone 16 under TS-2, Zones 2 and 6 under TS-5, and Zones 8 and 9 under TS-6. SVR's strong performance in these zones highlights its flexibility in handling diverse rainfall patterns. MLR model ranks next, with its best performance in three zones: Zones 1, 4, and 7, all under the TS-5 scenario. Although its scope is more limited compared to other algorithms, MLR demonstrates reliability in specific zones with consistently low prediction errors. On the other hand, the RFR algorithm only shows an advantage in one zone—Zone 14 under TS-1 and Zone 3 under TS-6. This suggests that while RFR performs well in certain conditions and zones, overall it is less competitive compared to ETR, SVR, or MLR in the tested scenarios.

Overall, the evaluation results for the TS scenarios indicate that eleven predictive models outperformed others, while the VR scenarios produced seven top-performing predictive models with high accuracy, as indicated by the low error values. The combination of algorithms and experimental scenarios in both approaches successfully developed predictive models with the best accuracy in each zone, indicated by the lowest MAE values. These models were then selected for further testing during the validation phase to assess their reliability under real-world conditions. The validation phase aims to evaluate how effectively the models can perform, particularly in addressing accuracy challenges, which are a primary concern at the XYZ Climatology Station. These results provide a solid foundation for implementing the best predictive models according to the characteristics of each zone, while also supporting efforts to improve rainfall forecast accuracy. The best predictive models selected for re-testing in the validation phase are summarized in Table 3.

Table 5. Summary of Selected Predictive Models by Scenario

Zone	Predictive Models for VR Scenarios	Predictive Models for TS Scenarios
1	MLR Model VR-3	MLR Model TS-5
2	SVR Model VR-1	SVR Model TS-5
3	SVR Model VR-1	ETR Model TS-6
4	MLR Model VR-3	MLR Model TS-5
5	MLR Model VR-2	SVR Model TS-1
6	SVR Model VR-2	SVR Model TS-5
7	ETR Model VR-1	MLR Model TS-5
8	MLR Model VR-2	SVR Model TS-6
9	ETR Model VR-1	SVR Model TS-6
10	ETR Model VR-1	ETR Model TS-5
11	ETR Model VR-2	ETR Model TS-2
12	ETR Model VR-1	ETR Model TS-4
13	SVR Model VR-3	ETR Model TS-4
14	ETR Model VR-1	RFR Model TS-1
15	ETR Model VR-1	ETR Model TS-1
16	SVR Model VR-2	SVR Model TS-2

Model Validation

In this stage, each of the selected predictive models was tested again to ensure their reliability and evaluate how effectively they can perform in meeting the accuracy targets that are a key focus of this study. The results showed that the selected predictive models, both for the VR and TS scenarios, significantly improved the accuracy of seasonal rainfall predictions compared to the initial performance of the XYZ Climatology Station. The detailed results of the model performance for both scenarios in predicting seasonal rainfall for the consecutive years 2022 and 2023 are presented in Table 4.

Table 6. Validation of Selected Predictive Model Performance for the Years 2022 and 2023

Zone	Predictive Models for VR Scenarios	Proportion of Correct (%) 2022	Proportion of Correct (%) 2023	Predictive Models for TS Scenarios	Proportion of Correct (%) 2022	Proportion of Correct (%) 2023
1	MLR Model VR-3	83.3	75.0	MLR Model TS-5	80.6	75.0
2	SVR Model VR-1	75.0	75.0	SVR Model TS-5	75.0	77.8
3	SVR Model VR-1	61.1	72.2	ETR Model TS-6	77.8	66.7
4	MLR Model VR-3	55.6	80.6	MLR Model TS-5	52.8	83.3
5	MLR Model VR-2	80.6	66.7	SVR Model TS-1	72.2	66.7
6	SVR Model VR-2	66.7	75.0	SVR Model TS-5	69.4	83.3
7	ETR Model VR-1	69.4	66.7	MLR Model TS-5	83.3	69.4
8	MLR Model VR-2	80.6	77.8	SVR Model TS-6	75.0	77.8
9	ETR Model VR-1	88.9	61.1	SVR Model TS-6	86.1	72.2
10	ETR Model VR-1	80.6	66.7	ETR Model TS-5	80.6	75.0
11	ETR Model VR-2	63.9	55.6	ETR Model TS-2	72.2	63.9
12	ETR Model VR-1	63.9	61.1	ETR Model TS-4	63.9	63.9
13	SVR Model VR-3	77.8	66.7	ETR Model TS-4	75.0	69.4
14	ETR Model VR-1	72.2	66.7	RFR Model TS-1	69.4	75.0
15	ETR Model VR-1	61.1	75.0	ETR Model TS-1	63.9	72.2
16	SVR Model VR-2	66.7	69.4	SVR Model TS-2	72.2	69.4
	Mean	**71.7%**	**69.4%**	**Mean**	**73.1%**	**72.6%**

The validation results for the predictive models selected in the VR and TS scenarios for 2022 and 2023, as shown in Table 4, indicate significant improvement compared to the initial performance of the XYZ Climatology Station. In the VR scenario, the average accuracy of the predictive model in 2022 reached 71.7%, an increase of 14.3% from the initial average accuracy of 57.4%. This achievement not only exceeded the organizational target of 70% but also demonstrated that the predictive model provided more accurate results than previous approaches. However, in 2023, the average accuracy of the VR model slightly decreased to 69.4%, falling below the organizational target, although still showing a 10.5% improvement from the initial accuracy of 58.8%. Meanwhile, the TS scenario provided more consistent results. In 2022, the average accuracy of the TS model reached 73.1%, an increase of 15.7% from the initial average of 57.4%, surpassing the organizational target. In 2023, although there was a slight decline, the average accuracy remained high at 72.6%, reflecting a 13.8% improvement from the initial accuracy of 58.8%. Overall, the validation results for both scenarios showed that the developed predictive models made a significant contribution to improving the accuracy of seasonal rainfall forecasts, with the TS scenario slightly outperforming in terms of average accuracy for both years.

In conclusion, both scenario approaches have successfully demonstrated the effectiveness of predictive models in substantially improving the accuracy of seasonal rainfall predictions. When comparing the two, the predictive models from the TS scenario showed an advantage in maintaining consistent accuracy over two consecutive years. This indicates that the TS scenario produced stable predictive models with good generalization ability in predicting the variability of seasonal rainfall across different years and climate conditions. This success also reflects the effectiveness of using sufficient historical data, providing a stronger foundation for capturing long-term patterns and generating more stable predictions. As a result, the TS scenario proves to be an excellent solution for implementing consistent and accurate seasonal rainfall forecasts.

D. Conclusion

This study aims to develop a machine learning-based predictive model that can improve the accuracy of seasonal rainfall forecasts across 16 climate zones. By utilizing five regression algorithms—MLR, SVR, ETR, RFR, and DTR—this research applies an ensemble approach to multiple model forecast data from various climate models. Experimental exploration was carried out through two main scenarios: variation in the number of input variables (VR) and the length of the time series data (TS), to determine the most accurate and reliable predictive model configuration. The results demonstrate that the time series-based model, with a longer data range, performs better in recognizing climate conditions and variability patterns, leading to more reliable and accurate seasonal rainfall predictions.Among the five regression algorithms tested in these experimental scenarios, four algorithms contributed to the improvement of prediction accuracy. SVR and ETR emerged as dominant, with accuracy improvements in 11 out of 16 zones, or approximately 69% of the areas. These findings indicate that both algorithms are effective in identifying patterns from historical climate model outputs and observed rainfall data in the 16 zones, which could potentially be used operationally to enhance the reliability of seasonal rainfall forecasts in the region under the responsibility of Station XYZ.

E. Acknowledgment

I would like to express my sincere gratitude to the climatology station XYZ, especially the operational forecaster staff, for their invaluable guidance, support, and the provision of data during the development of this study. Furthermore, I wish to extend my thanks to the Education and Training Center of the Meteorology, Climatology, and Geophysics Agency (BMKG) for their financial support, which has played a crucial role in enabling the completion of this academic work and the publication of this study.

F. References

[1] Badan Meteorologi Klimatologi dan Geofisika (BMKG), *Perkiraan Musim Hujan 2023/2023*, no. September. Jakarta: Kedeputian Bidang Klimatologi BMKG, 2023.

[2] R. Indonesia, *Undang-Undang Republik Indonesia Nomor 31 Tahun 2009 Tentang Meteorologi, Klimatologi, Dan Geofisika*. Indonesia, 2009, pp. 1–60.

[3] B. M. K. dan G. Stasiun Klimatologi XYZ - BMKG, *Perjanjian Kinerja Tahun 2022/2023*. Indonesia, 2022, p. 1.

[4] C. I. Arrijal Rachman, "La Nina 3 Tahun Berakhir, Ancaman Baru Intai RI di 2023." Accessed: Oct. 13, 2024. [Online]. Available: https://www.cnbcindonesia.com/news/20221229190500-4-401309/la-nina-3-tahun-berakhir-ancaman-baru-intai-ri-di-2023

[5] V. I. Anugrah Andriansyah, "BMKG: Fenomena El Nino Bertahan hingga Akhir Tahun 2023." Accessed: Oct. 13, 2024. [Online]. Available: https://www.voaindonesia.com/a/bmkg-fenomena-el-nino-bertahan-hingga-akhir-tahun-2023/7204967.html

[6] G. Flato *et al.*, "Evaluation of climate models," in *Climate change 2013: the*

physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2014, pp. 741–866.

[7] M. Collins *et al.*, “Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles,” *Clim. Dyn.*, vol. 36, pp. 1737–1766, 2011.

[8] Y.-M. Min, V. N. Kryjov, S. M. Oh, and H.-J. Lee, “Skill of real-time operational forecasts with the APCC multi-model ensemble prediction system during the period 2008–2015,” *Clim. Dyn.*, vol. 49, no. 11, pp. 4141–4156, 2017.

[9] H. Xu and R. B. Street, “APEC Climate Symposium 2012,” 2012.

[10] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.

[11] A. Géron and R. Russell, “Machine learning step-by-step guide to implement machine learning algorithms with Python,” *O’Reilly Media, Inc*, p. 106, 2019.

[12] A. Sani, “Machine Learning for Decision Making.” Université de Lille 1, 2015.

[13] A. Sharma, A. Khanna, M. Bhargava, and R. Pendse, “Rainfall Prediction: analysis of machine learning algorithms and ensemble techniques,” in *2021 7th International Conference on Signal Processing and Communication (ICSC)*, IEEE, 2021, pp. 234–240.

[14] T. Li, Z. Jiang, H. Le Treut, L. Li, L. Zhao, and L. Ge, “Machine learning to optimize climate projection over China with multi-model ensemble simulations,” *Environ. Res. Lett.*, vol. 16, no. 9, 2021, doi: 10.1088/1748-9326/ac1d0c.

[15] G. K. Uyanık and N. Güler, “A study on multiple linear regression analysis,” *Procedia-Social Behav. Sci.*, vol. 106, pp. 234–240, 2013.

[16] B. Pang, J. Yue, G. Zhao, and Z. Xu, “Statistical downscaling of temperature with the random forest model,” *Adv. Meteorol.*, vol. 2017, no. 1, p. 7265178, 2017.

[17] M. J. Themeßl, A. Gobiet, and A. Leuprecht, “Empirical-statistical downscaling and error correction of daily precipitation from regional climate models,” *Int. J. Climatol.*, vol. 31, no. 10, pp. 1530–1544, 2011.

[18] K. Ahmed, D. A. Sachindra, S. Shahid, Z. Iqbal, N. Nawaz, and N. Khan, “Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms,” *Atmos. Res.*, vol. 236, no. July 2019, p. 104806, 2020, doi: 10.1016/j.atmosres.2019.104806.

[19] P. C. Deka, “Support vector machine applications in the field of hydrology: a review,” *Appl. Soft Comput.*, vol. 19, pp. 372–386, 2014.

[20] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, pp. 3–42, 2006.

[21] R. Xu, N. Chen, Y. Chen, and Z. Chen, “Downscaling and projection of multi-CMIP5 precipitation using machine learning methods in the upper Han River basin,” *Adv. Meteorol.*, vol. 2020, no. 1, p. 8680436, 2020.

[22] L. Xu, N. Chen, X. Zhang, Z. Chen, C. Hu, and C. Wang, “Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning,” *Clim. Dyn.*, vol. 53, pp. 601–615, 2019.

[23] L. Breiman and A. Cutler, “Setting up, using, and understanding random

forests V4. 0," *Univ. California, Dep. Stat.*, 2003.
[24] L. Breiman, *Classification and regression trees*. Routledge, 2017.
[25] M. Mendez, B. Maathuis, D. Hein-Griggs, and L.-F. Alvarado-Gamboa, "Performance evaluation of bias correction methods for climate change monthly precipitation projections over Costa Rica," *Water*, vol. 12, no. 2, p. 482, 2020.
[26] H. A. Bhatti, T. Rientjes, A. T. Haile, E. Habib, and W. Verhoef, "Evaluation of bias correction method for satellite-based rainfall data," *Sensors*, vol. 16, no. 6, p. 884, 2016.
[27] I. T. Jolliffe and D. B. Stephenson, *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons, 2012.
[28] R. Muharsyah, "Kajian verifikasi produk prakiraan curah hujan bulanan (2003-2012)," *J. Meteorol. Dan Geofis.*, vol. 18, no. 1, 2017.