ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Long-Term Seasonal Rainfall Forecasting Using Regression Analysis and Artificial Neural Network with Larg-Scale Circulation Indices

Ketvara Sittichok[1*], Napatsorn Rattanapan[2], and Rittisak Sakulkaew[3]

[1] Research Center for Sustainable Development, Faculty of Engineering at Kamphaengsaen Campus, Kasetsart University, Nakhon Pathom, 73140, Thailand; fengkrs@ku.ac.th
[2] Faculty of Engineering at Kamphaengsaen Campus, Kasetsart University, Nakhon Pathom, 73140, Thailand; napatsorn.rtnp@gmail.com
[3] Faculty of Engineering at Kamphaengsaen Campus, Kasetsart Univers, Nakhon Pathom, 73140, Thailand; rittisak.vpz@gmail.com
* Correspondence: E-mail: fengkrs@ku.ac.th

Citation:
Sittichok, K.; Rattanapan, N.; Sakulkaew, R. Long-term seasonal rainfall forecasting using regression analysis and artificial neural network with large-scale circulation indices. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(3), e252507. https://doi.org/10.55164/ajstr.v27i3.252507.

Article history:
Received: April 2, 2024
Revised: April 26, 2024
Accepted: April 27, 2024
Available online: April 30, 2024

Abstract: Several months in advance, long-term rainfall prediction plays an important role in water management, especially for countries dependent on agriculture. The objective of this study is to forecast the long-term rainfall of eight rain gauge stations in the Phetchaburi River Basin, Thailand, 12–18 months in advance using linear regression (simple linear regression (SLR)/multiple linear regression (MLR)) and non-linear relations (polynomial regression (PR)/artificial neural network (ANN)). Seven atmospheric circulation indices, ONI, DMI, MEI V. 2, NINO4, NINO3.4, NINO3, and NINO1+2, and historical rainfall data were used as predictors in the models. To avoid bias in empirical equation construction, one-year cross-validation was also applied together with a one-month moving window average approach from January to July of the preceding year (12–18 months lead time) to seek suitable periods of predictors for predicting rainfall. The results reveal that the surface temperature indices of the Indian Ocean (DMI) and Pacific Ocean (NINO) are the most essential for forecasting rainfall. MEIV2 and ONI were only positively correlated with local rainfall when non-linear regression was used. Non-linearity models showed better forecasting skills compared to linear regression. The suitability of periods varied according to the statistical models and selected predictors.

Keywords: Rainfall forecasting; Teleconnection; Statistical forecasting; Atmospheric Circulation Indices

1. Introduction

The impacts of climate change on various issues can now be seen in many regions of the globe, especially rainfall variability. Changes in increasing and decreasing rainfall, including alterations in the rainfall pattern, significantly affect people in different areas [1]. Rainfall also plays an important role in the hydrological cycle since streamflow and groundwater affect many activities, such as agriculture, livestock, and the water supply. Changes in pattern and the amount of rainfall lead to challenges in effective water management. Therefore, the ability to predict rainfall is important to government agencies in planning water allocation and management strategies to avoid water scarcity and flood damage.

Several months in advance, long-term rainfall prediction is essential for effective water management, especially in countries heavily dependent on agricultural activities. In addition, rainfall prediction is crucial for mitigating the effects of extreme flood and drought events. Therefore, an effective rainfall forecasting method should be developed to protect and improve human lives and the environment. As part of a complex atmospheric phenomenon, rainfall varies from one region to another. Several large-scale circulation indices are proven to be significantly related to rainfall [2, 3]. El Nino Southern Oscillation (ENSO), commonly used for rainfall forecasting, significantly influences seasonal precipitation across the globe [4]. [5] investigated the relationship between ENSO signals and rainfall, finding that they showed statistical significance and were suitable for predicting rainfall in Sri Lanka. ENSO and the Indian Ocean Dipole (IOD) were also applied for long-term seasonal rainfall forecasting in Australia in the research work conducted by [6]. Various research works have also proven that El Nino and La Nina signals strongly influence local precipitation [7]. Other climate indices applied to seek a relationship with local rainfall for effective rainfall forecasting consist of the Oceanic Nino Index (ONI), Multivariate ENSO Index (MEI), Dipole Mode Index (DMI), Nino1+2, Nino3, Nino 4, and Nino 3.4. These climatic indices have been widely used in different regions to forecast precipitation [8, 9]. Nino 3.4 also performed well as a statistical method for forecasting rainfall in Bandung, Indonesia [10].

Forecasting models can be divided into dynamical and statistical. Various equations relating to the individual characteristics of the atmosphere, ocean, and land and their relations are employed in dynamical models. These models can generate weekly, monthly, or seasonal rainfall forecasts. Still, long-term forecasting carried out more than several months in advance exhibits high uncertainty since the initial conditions and a temporal framework must be determined. On the other hand, statistical models are widely used because of their capability to provide a longer forecasting lead time [11]. Various statistical techniques for rainfall forecasting have been proposed in large regions all over the world. Linear and non-linear regression are commonly employed where long-term rainfall forecasting is required. Kim C. G. et al. [11] used multiple linear regression (MLR) with lagged correlation to forecast monthly precipitation 12 months in advance with acceptable results over long periods. Six statistical methods: MLR, Multi-layer Perceptron, Pace regression, Radial Basis Function, K-star Algorithm, and Bootstrap Aggregating (bagging) were employed by Gnanasankaran, N., et al. [12] to generate rainfall forecasts, and the results showed MLR to be the most effective method for forecasting rainfall in this area. [13] forecasted rainfall in tropical regions in a seasonal time scale using MR and polynomial regression (PR), revealing that the model provided moderate to good forecasting results for long-term rainfall over 5–12 months.

Regression analysis and the Artificial Neural Network (ANN) have attracted the attention of researchers developing statistical forecasting models with long lead times. The ANN is a machine learning method that can predict non-linear relationships between various large-scale climate indices and rainfall. [14] attempted to predict rainfall for 3–12 months using the ANN in Eastern Australia, revealing that better performance was achieved with more extended historical data and single-month optimization. Monthly rainfall forecasts were also studied by Lee, J. et al. [15] using the ANN with climate indices in Korea. They concluded that the ANN and Monte-Carlo cross-validation could provide acceptable medium-term rainfall forecasting. Works on rainfall forecasting on a monthly and seasonal time scale over the medium and long term can be found in [16-18].

This study uses statistical methods, large-scale climate indices, and suitable periods for rainfall forecasting 12–18 months ahead in the Phetchaburi River Basin, Thailand. Seven circulation indices and the individual historical rain for each station were forced into statistical models with a moving average window of a one-month lag time. Statistical models comprising regression analysis (SLR, MLR, and PR) and the ANN were employed to search for a suitable method at each station. However, the model bias from using the same dataset for constructing empirical equations and the validation process needed to be considered, especially for the statistical model. The one-year cross-validation method was subsequently applied in this research. Three objective functions were used to evaluate the model, namely the correlation coefficient (R), root mean square error (RMSE), and percent bias (PBIAS).

2. Materials and Methods

2.1 Study area

The Phetchaburi River Basin is a 6,255 km^2 watershed in the middle of the Kingdom of Thailand (**Figure 1**). Three major tributaries (the Bangkloy, Maepradone Maeprchan, and Huaypak Rivers) drain into the main river (the Phetchaburi River) and play a vital role in people's livelihoods by providing water for agriculture, mostly rice cultivation. There are 224 water resource development projects in the river basin, the most important being the Kaeng Krachan and Mae Prachan reservoirs, which deliver water for agricultural, domestic, and other uses. The mean annual precipitation in the watershed is around 1,110 mm, concentrated in the primary rainy season, which spans from May to October, with minimum, average, and maximum around 200, 500, and 900 mm, respectively. The discharge of the basin follows the same pattern as the rainfall and is around 300–600 mcm during the rainy season. The average temperature is around 28 ^{0}C, peaking in April and a low in December. Both floods and droughts tend to occur at regular intervals in the basin. The lower sub-basin was the primary location where around 197 houses were impacted by the severe floods reported in 2016, 2017, and 2018.

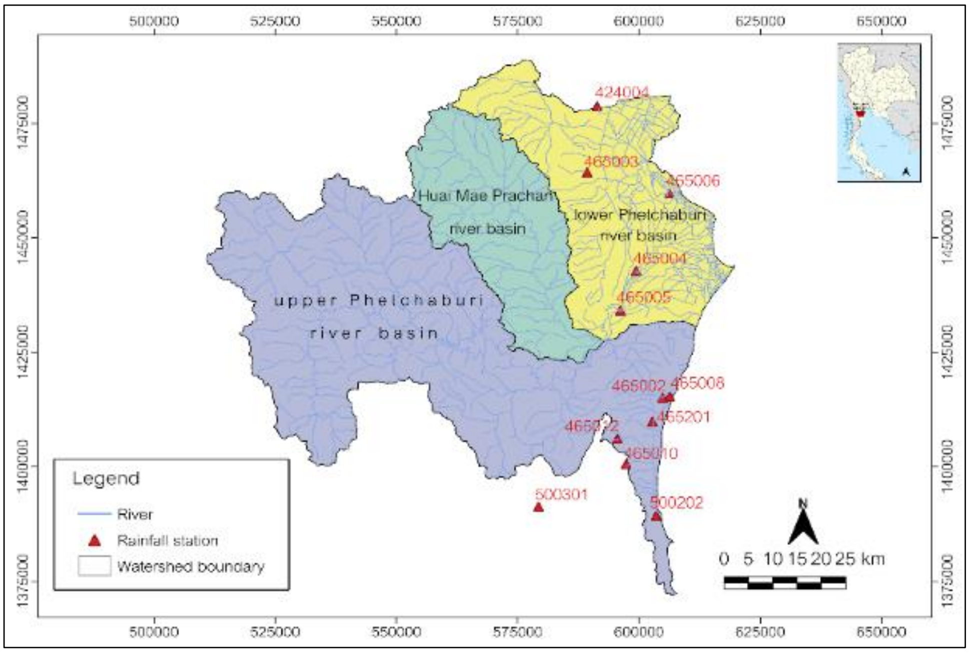

Figure 1. Phetchaburi River Basin and rainfall stations.

2.2 Data

This study uses two types of datasets: large-scale circulation indices and the observed rainfall in the basin. Details of both datasets are presented as follows.

2.2.1 Atmospheric Circulation Index

Seven atmospheric circulation indices, ONI, DMI, MEI V.2, NINO 1+2, NINO 3, NINO 3.4, and NINO 4 over 36 years (1985–2020) are used as predictors. The variability of these predictors is shown in **Figure 2**

- **Oceanic Nino Index (ONI):** The ONI is calculated from the change in sea surface temperature, its value defining an El Nino event as being equal to or higher than +0.5 ^{0}C, whereas an event equal to or less than -0.5 ^{0}C indicates the occurrence of the La Niña phenomenon. The National Oceanic and Atmospheric Administration (NOAA) uses the ONI to predict the occurrence of ENSO and analyze its severity. It is considered weak, moderate, or strong within the plus or minus 0.5–0.9, plus or minus 1.0–1.4, and greater than or equal to 1.5, respectively (The Meteorological Department of the Marine Meteorological Center, 2012). The monthly average ONI from 1985–2020 is presented in **Figure 2**.
- **Dipole Mode Index (DMI):** This is used to measure the condition of the Indian Ocean Dipole (IOD), a phenomenon occurring around the equator in the Indian Ocean. The difference in sea surface temperature (SST) between the west and east coasts of the Indian Ocean is considered. A positive IOD is determined when the west coast SST of the Indian Ocean is noticeably cooler and the east coast SST is noticeably warmer and vice versa. A positive/negative IOD leads to less/more rainfall in Thailand. **Figure 2** shows the average monthly data according to the DMI from 1985–2020 (Marine Meteorological Center Meteorological Department, 2012).

- **Multivariate ENSO Index Version 2 (MEI V.2)**: This combines five large circulation factors: sea level pressure, SST, zonal wind component, meridional wind component, and outgoing longwave radiation. A negative MEI portrays the chance of the La Niña phenomenon occurring. Meanwhile, if the MEI value is positive, it shows the chance of the El Niño phenomenon occurring (National Oceanic and Atmospheric Administration, 2021). **Figure 2** shows the average monthly data for the DMI from 1985–2020.
- **NINO SST Indices:** NINO estimates the SST change in the central Pacific Ocean. The measurement of the change in SST is divided into four levels: NINO 1+2 (0–10S, 90W - 80W), NINO 3 (5N–5S, 150W–90W), NINO 3.4 (5N–5S, 170W–120W), and NINO 4 (5N–5S, 160E–150W) (National Center for Atmospheric Research, 2018) (Figure 2). All these indices measure the severity of La Niña and El Niño phenomena. **Figure 2** presents the average monthly data for NINO 1+2, NINO 3, NINO 3.4, and NINO 4 from 1985–2020.

2.2.2 Rainfall

8 Daily rainfall measurements showing completed data prepared by the Meteorological Department are used as both predictor and predictor for the statistical models in this study. The amount of seasonal rainfall during the rainy season from August to October during the chosen period of the year is used as the predictor. Meanwhile, the forecasted amount of rainfall before the rainy season (12−18 months lead time) is used together with the atmospheric circulation index as the predictor. **Table 1** shows the details and locations of the rainfall measurement stations.

2.3 Statistical methods

2.3.1 Regression analysis

Various statistical methods have been applied in this study. Three regression methods, SLR, MLR, and PR, were used to construct the linear and non-linear equations (Equations 1–3) and obtain the rainfall forecasts. SLR and MLR were used for linear regression analysis with one and more than one predictor, respectively. PR was employed as a non-linear regression method at a power level of 2–5. The moving average method was used to predict rainfall 12–18 months in advance for all predictors. Monthly data on each predictor were collected, starting from January, one year earlier than the forecasted year ($x_{jan(Y-1)}$) to July of the same year ($x_{jul(Y-1)}$), to calculate the moving average and develop a specific period for the predictor. The starting month was also changed from January to July. The example predictor for each lead time is indicated in Table 2. Therefore, 28 periods were tested in each statistical model.

$$y = a + bx \tag{1}$$

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_i \tag{2}$$

$$y = a + b_1x_1 + b_2x_2^2 + b_2x_3^3 + \cdots + b_nx_i^n \tag{3}$$

Model bias should be considered when employing statistical forecasting models to avoid overestimating the results. Since the training data used to construct the relationship between predictors and predictand should not be applied for testing, one-year cross-validation was performed in this study, and the hindcast between 1986 and 2020 was investigated. Each forecasted year was left out of the model construction step, and all rainfall forecasts were then compared to the observations.

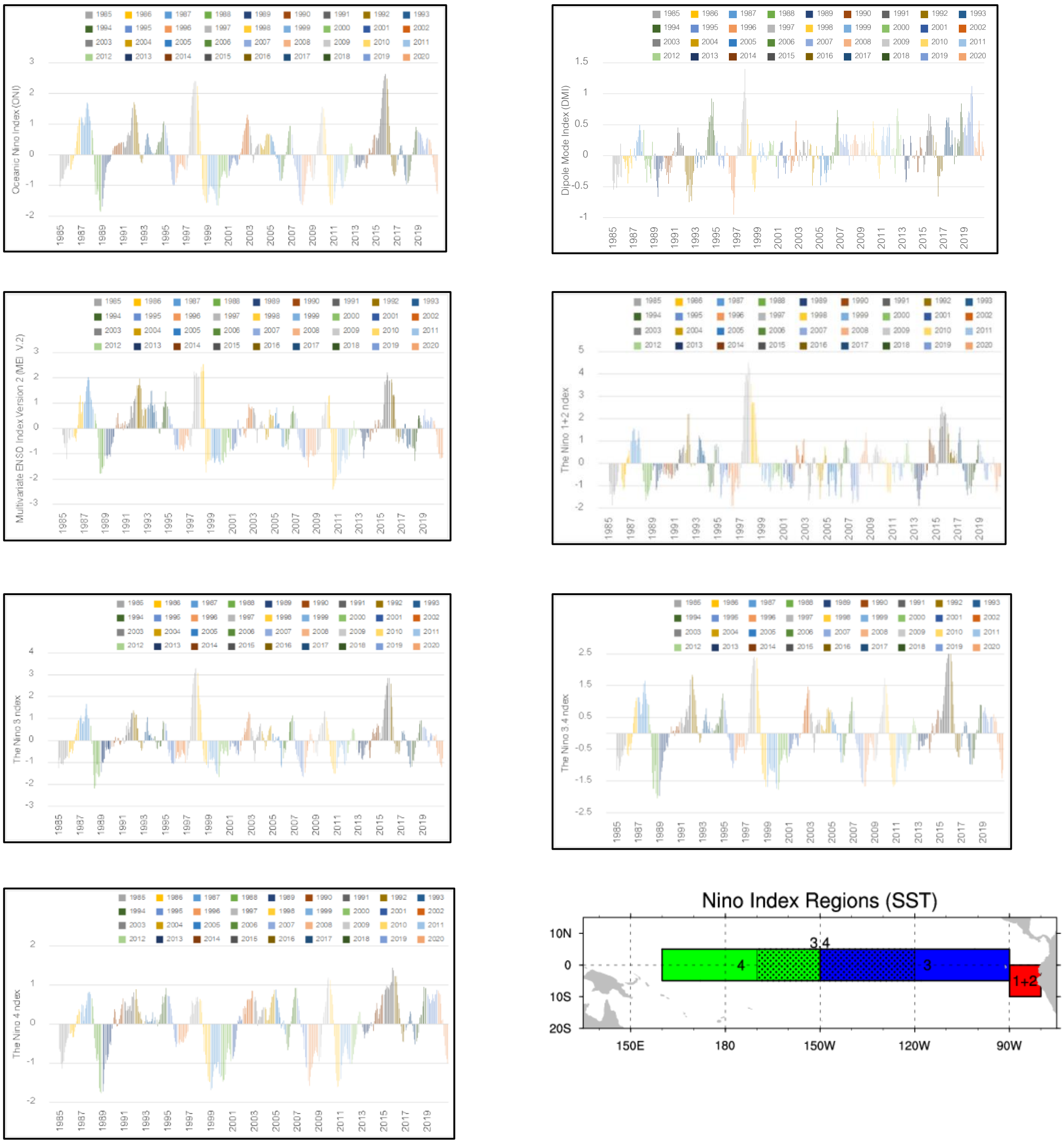

Figure 2. Average monthly data from the atmospheric circulation indices from 1985 to 2020

Table 1. Details of rain gauge stations

Station number	Station name	Latitude/ Longitude
465002	Cha-am Agricultural Office	12°47'59.6"N / 99°58'00.1"E
465004	Ban lad Agricultural Office	13°02'56.7"N /99°56'10.3"E
465008	Cha-am Forest Training Center	12°48'03.6"N/99°58'56.3"E
465010	Forest Development Project	12°41'52.4"N/99°54'09.1"E
465012	Somdet Phrasrinagarindra Park	12°43'24.2"N/99°53'36.0"E
465201	Phetchaburi Weather Station	12°59'58.8"N/100°03'39.7"E
500202	Hua-Hin Weather Station	12°34'42.6"N/99°57'17.1"E
500301	Nong-Pub Weather Station	12°35'21.0"N/99°44'04.3"E

2.3.2 Artificial neural networks

The artificial neural network (ANN) comprises a parallel distributed processor with many processing units inside the structure, initially inspired by the human brain, which contains neurons to process information. Many inputs can be received through this process, and relevant outputs can be generated. The main ANN process comprises a set of connecting links, an activation function, and bias [19]. Liu, Q. et al. [20] outlined the strengths of ANNs for rainfall forecasting. They mentioned that the ANN is a parallel process, effectively working on a large amount of data. It is also a data-driven model that is not subject to modeling restrictions. It can collect many experiences to learn how to predict rainfall patterns with complicated non-linear relationships between datasets. Equations 4–5 present the mathematical expression of ANN comprising neuron pre-activation or input activation (Equation 4) and neuron (output) activation (Equation 5), whereas x_i is the input data i, W the connection weights, and b and g the neuron bias and activation function, respectively.

$$\mathrm{a(x)} = b + \sum_i w_i x_i = b + W^T x \tag{4}$$

$$h(x) = g\big(a(x)\big) = g(b + \textstyle\sum_i w_i x_i) \tag{5}$$

ANN was also applied for rainfall forecasting using a similar process to regression, consisting of the cross-validation step and a one-month moving average. Since high efficiency is essential when dealing with many input variables, all predictors were forced into the models.

Table 2. Examples of predictor calculations with a one-month window

Start month	End month	Forecasted lead time	Predictor calculation
Jan	Jan	18	$x = x_{jan(Y-1)}$
Jan	Feb	17	$x = \frac{(x_{jan(Y-1)} + x_{feb(Y-1)})}{n}$
Jan	…		
Jan	July	12	$x = \frac{(x_{jan(Y-1)} + x_{feb(Y-1)} + x_{...(Y-1)} + x_{jul(Y-1)})}{n}$
Feb	Feb	17	$x = x_{feb(Y-1)}$
Feb	Mar	16	$= \frac{(x_{feb(Y-1)} + x_{mar(Y-1)})}{n}$
Feb	…		
Feb	July	12	$x = \frac{(x_{feb(Y-1)} + x_{mar(Y-1)} + x_{...(Y-1)} + x_{jul(Y-1)})}{n}$
…	…		
July	July	12	$x = x_{jul(Y-1)}$

2.3.3 Model estimation

Three objective functions were used to estimate the forecasted results in this study: the correlation coefficient (R), root mean square error (RMSE), and percentage of bias (PBIAS), as presented in Equations 6–8, respectively. y_i and p_i are the measured and predicted values of the forecasted rainfall in a year i. The r value ranges from -1 to 1, and only positive values should be considered when estimating the results. The RMSE ranges from 0 - ∞ and the predicted rainfall is correlated with the measured rainfall if the RMSE value is close to 0. Finally, a PBIAS of 0 represents the best performance of the model, whereas a higher PBIAS indicates an inferior performance. Unfortunately, no explicit criteria explain the acceptable outcomes of long-lead-time rainfall forecasting. Most research studies reported model estimates with a range of R values from 0.3-0.8, which varied based on factors such as the methodology used, the study area, and the predictors considered, as shown in Hossain, Kim, Singaroodi, et al. [2, 6, 11, 21].

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(p_i - \bar{p})}{(\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2})\,(\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2})} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n} \times \sum (y_i - p_i)^2} \tag{7}$$

$$PBIAS = [\frac{\sum_{i=1}^{n}(p_i - y_i)}{\sum_{i=1}^{n}(y_i)}] \times 100 \tag{8}$$

2.4 Research methodology

Statistical models strongly rely on the quality of predictors and predictors; therefore, all input data require initial verification. All missing values were completed using the inverse distance weight method. The 28 independent variables of each predictor calculated from each period with the moving average method were then prepared and tested to seek the best period for forecasting using SLR. The most effective predictor of a suitable period using the SLR at each station was then reported. However, using a combination of predictors may lead to better forecasts, and MLR was then performed in this step. Only predictors demonstrating a relationship with rainfall at each station in the SLR model were selected for this second process. The number of models at each station depends on the selected predictors. Finally, PLR was employed for a non-linear regression test at 2–5 degrees. Cross-validation and moving averages were still used in these steps. All forecasting results were investigated to ascertain their relationship with observed rainfall using the three objective functions mentioned in the previous section. The effective predictors of high performance for forecasting rainfall in this area were then reported. An overview of this research is presented in **Figure 3**.

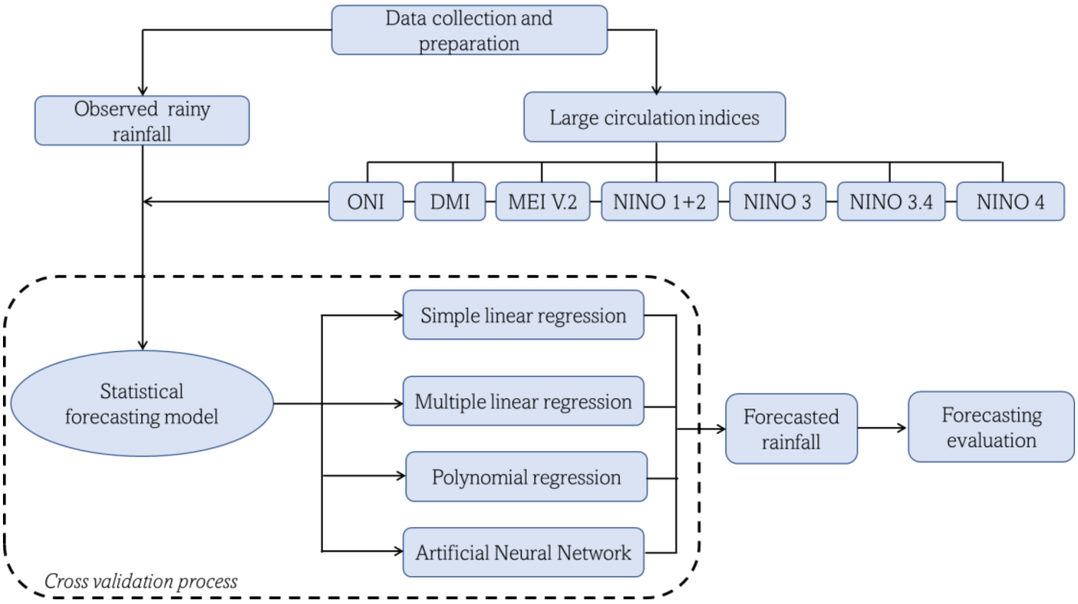

Figure 3. Methodological process

3. Results and Discussion

Teleconnection between each large circulation index and the observed rainfall for 12 stations was analyzed using SLR. Autocorrelation with a moving average window of one-month lag time was applied to search for the best period. A combination of large circulation indices and previous rainfall predictors was then used to MLR, PR, and ANN. Suitable predictors and periods are finally informed.

3.1 Forecasted rainfall using regression analysis

3.1.1 Simple linear regression

The relationship between each predictor and local rainfall was initially examined. Climate indices and historical rainfall were used as predictors in SLR, and rainfall forecasts for the rainy season in the basin for the period from 1985 to 2020 were generated. **Table 3** indicates the predictors showing a positive relationship with observed rainfall. The results revealed that the phenomenon around the Indian Ocean (DMI) effectively predicted most (seven) stations. Historical rainfall at the individual stations and NINO 1+2 were also considered effective in forecasting long lead times in this area. MEI V2 and NINO 3.4 variables could not capture changes in local rainfall, whereas NINO 3 and NINO 4 were suitable to be applied as input variables for SLR only at one station.

Table 3. Effective predictors used in SLR for each station

Station	ONI	DMI	MEI V.2	NINO 1+2	NINO 3	NINO 3.4	NINO 4	Historical rainfall	Range of R
465002		×		×					0.30–0.32
465004		×							0.37
465008		×						×	0.28–0.37
465010		×		×				×	0.21–0.34
465012		×							0.27
465201					×				0.26
500202		×						×	0.33–0.48
500301		×							0.34

Table 4 presents the best period and rainfall predictor using SLR 12–18 months in advance. Forecasting estimations using R, RMSE, and PE to obtain the best period and predictor for each station are shown in the table. All predictors start from January–July of the previous year (Y-1) with a one-month lag time and moving average window forced into the model. The results revealed that DMI was the most influential predictor for seven stations, demonstrating R values between 0.27 and 0.51, while the RMSE and PBIAS ranged from 132.2 to 179.0 mm and 13.9 to 30.4%, respectively. Only one station (465201) showed NINO 3 as the best predictor, giving an R, RMSE, and PBIAS of 0.25, 139.0 mm, and 18.8%, respectively. The most suitable prediction periods for forecasting were different at each station. The best rainfall forecast could be observed at ST-50020 with an R of 0.51, RMSE of 139.4, and PE of 20.0%. **Figure 4** indicates the rainfall in the wet season from 1992–2020. As can be observed from the SLR forecasting results, it could not capture extreme events for all stations. For example, the observed rainfall in 1999 and 2003 at ST-465002 increased to 752 and 835 mm, while forecasted rainfall amounts of 400 and 458 mm were generated using SLR. This was similar to the SLR rainfall forecasts at ST-500202, which could not reach these extreme events with observed/forecasted rainfall amounts of 725/429 and 768/443 mm. However, the models could generate reliable forecasting results during average rainfall circumstances.

3.1.2 Multiple linear regression

All predictors were combined to investigate the model efficiency using MLR. 255 models were tested for each station, each examined to find the most suitable prediction period 12–18 months in advance. These models used a combination of large circulation indices and historical rainfall, providing better rainfall forecasting skills at most stations. **Table 5** presents details of the most efficient predictors and the number of models providing R values higher than 0.3. The most significant combination was ST-465004, with 98 models using all predictors. Like the SLR model, DMI and historical rainfall remained the most effective predictors of long-lead time rainfall. NINO3.4 was the second-best predictor, showing good results when employed with other predictors (four stations). This result differed significantly from the SLR forecasting process. It is also worth considering that some variables presented higher prediction efficiency when used with others, such as NINO 3, NINO 1+2, and NINO 4.

The R, RMSE, and PBIAS ranges for all stations were 0.32–0.57, 132.6–176.9 mm, and 13.7–31.0%, respectively (**Table 4**). The best forecasting results could be found at ST-465004, which has the most significant number of predictors (six). A few extreme events in historical periods were captured by MLR forecasting (**Figure 4**). The number of rainfalls in 1993 at ST-465002 was 776 mm, differing slightly from the MLR forecast of 725 mm. In 1999, the rainy season and MLR forecasting results for ST-465010 were 682 and 706, respectively.

Table 4. Statistical forecasting results for the rainy season

Station	Method	Predictor	Period	R	RMSE (mm)	PBIAS (%)
465002	SLR	DMI	Apr-May	0.41	155.5	25.9
	MLR	DMI, NINO1+2	Apr–May	0.52	147.5	24.4
	PR	ONI	Feb	0.53	153.4	26.5
	ANN	(all)	Apr	0.42	162.9	28.9
465004	SLR	DMI	May	0.35	149.3	25.8
	MLR	ONI, DMI, MEI V.2, NINO 4, NINO 3.4, NINO 3	Apr–May	0.57	137.1	21.6
	PR	MEI V.2	Mar–Apr	0.49	142.6	24.7
	ANN	(all)	April–May	0.52	144.2	22.5
465008	SLR	DMI	Jan–Jun	0.40	145.6	30.4
	MLR	DMI, Rainfall	May–Jun	0.37	150.4	31.0
	PR	DMI	May–Jul	0.41	148.2	32.7
	ANN	(all)	Mar–Jul	0.41	146.9	28.8
465010	SLR	DMI	May–Jul	0.34	142.5	25.6
	MLR	NINO 1+2, Rainfall	Jan–Apr	0.35	142.9	27.1
	PR	MEI V2	Apr–Jul	0.43	148.1	28.3
	ANN	(all)	Jan–Jul	0.50	131.0	24.6
465012	SLR	DMI	May	0.27	179.0	26.8
	MLR	ONI, NINO 3.4	Apr	0.32	176.9	30.6
	PR	NINO 4	Jun	0.37	178.4	29.8
	ANN	(all)	Feb	0.59	150.4	25.8
465201	SLR	NINO 3	Jun	0.25	139.0	18.8
	MLR	DMI, NINO 3.4, NINO 3, Rainfall	Jun	0.38	135.1	19.6
	PR	NINO3	Jul	0.38	136.1	20.8
	ANN	(all)	May–Jul	0.40	131.8	17.8
500202	SLR	DMI	May	0.51	139.4	20.0
	MLR	DMI, NINO 4	May	0.49	142.1	20.5
	PR	DMI	May	0.42	183.8	29.7
	ANN	(all)	May–Jul	0.46	144.9	23.1
500301	SLR	DMI	May	0.37	132.2	13.9
	MLR	ONI, DMI, NINO 4, NINO 3.4, NINO 1+2	Apr–May	0.37	132.6	13.7
	PR	NINO 3.4	Jan–Feb	0.33	136.0	15.6
	ANN	(all)	May	0.58	116.7	13.2

Table 5. Effective predictors used in MLR for each station

Station	Models	ONI	DM I	MEI V.2	NINO 1+2	NIN O 3	NINO 3.4	NINO 4	Historica l rainfall	R Range
465002	7		x		x	x			x	0.32–0.52
465004	98	x	x	x	x	x	x	x	x	0.30–0.57
465008	7		x						x	0.30–0.37
465010	4		x		x				x	0.30–0.35
465012	3	x	x				x		x	0.30–0.32
465201	5		x			x	x		x	0.30–0.38
500202	4		x					x	x	0.30–0.49
500301	33	x	x		x		x	x	x	0.30–0.37

3.1.3 *Polynomial linear regression*

Another method used to forecast rainfall in this study is polynomial regression analysis (PR). Degrees of power ranging from 2–5 were applied to examine the association between all predictors and local rainfall in a non-linear regression area. Highly skilled predictors showing a positive correlation between observed and forecasted rainfall are presented in **Table 6**. Observed rainfall was used to predict long-lead time seasonal rainfall at each station, followed by DMI and ONI, respectively. The first two predictors (historical rainfall and DMI) showed similar results to when MLR was used for forecasting. However, the use of ONI and MEI V.2 showed a noticeable difference between these two methods. MEI V.2 only generated corresponding trend patterns for one station using MLR, whereas positive trends could be seen at four stations using PR. A similar situation arose with ONI at three stations using the MLR method, increasing to six stations for PR.

Tabled 6. Effective predictors used in PR for each station

Station	ONI	DMI	MEI V.2	NINO 1+2	NINO 3	NINO 3.4	NINO 4	Historical rainfall	R Range
465002	x	x				x		x	0.32–0.53
465004			x		x	x	x	x	0.32–0.49
465008	x	x						x	0.31–0.41
465010	x	x	x	x	x	x	x	x	0.33–0.43
465012	x	x		x	x		x	x	0.30–0.37
465201		x			x		x	x	0.34–0.38
500202	x	x	x		x	x		x	0.40–0.42
500301	x	x	x			x		x	0.31–0.33

3.2. *Forecasted rainfall using ANN*

The ANN is another statistical forecasting method employed to identify non-linear relationships between predictors and predictands. In similarity to regression analysis, all eight predictors with different one-month moving average periods were forced into the ANN application. The rainfall forecast for all stations showed R values equal to or higher than 0.40. The largest R value of 0.59 occurred at ST-465012. Minimum/maximum values of RMSE and PBIAS were 116.7 mm and 15.6%/162.9 mm and 32.7%, respectively. Five out of eight stations presented better performance using the ANN compared to all regression methods. The best overall statistical index could be found at ST-500301 with an R, RMSE, and PE of 0.58, 116.7, and 13.2%, respectively. However, when extreme rainfall events were considered, the ANN produced an inferior performance for rainfall forecasting compared to polynomial regression. The ANN showed good prediction ability for seasonal rainfall occurrence in 1992 at ST-465004, with observed and forecasted rainfall being 737 and 791 mm, respectively (**Figure 5**).

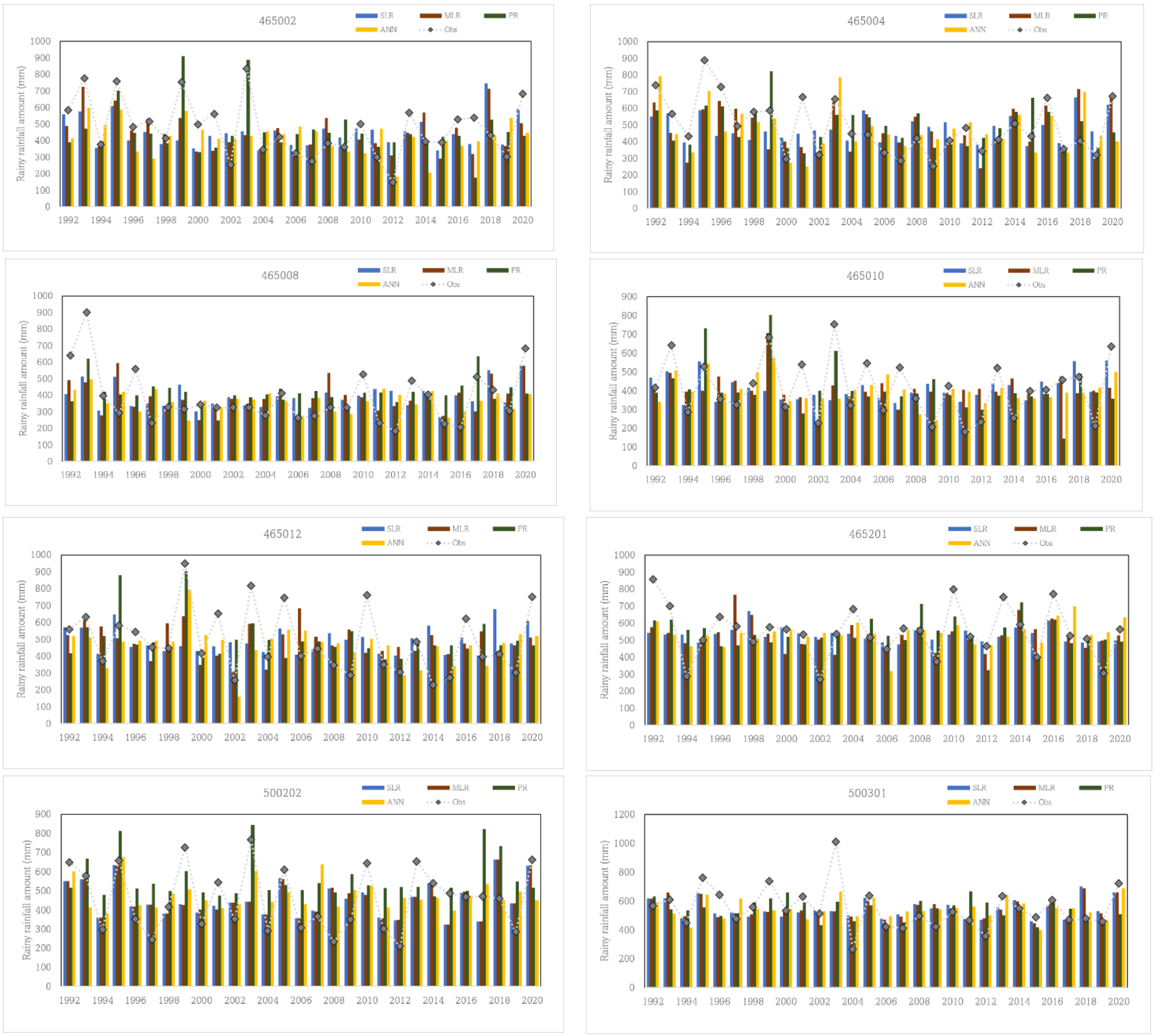

Figure 4. Observed and forecasted amounts of rainfall in the rainy season

3.3 Comparison between statistical methods for seasonal rainfall forecasting

Three regression approaches and the ANN were employed to forecast rainfall for eight stations 12–18 months in advance using large-scale circulation indices as predictors. The differences in sea surface temperatures between the west and east coasts of the Indian Ocean influenced rainfall in the study area. Seven SLR and PR stations and eight MLR stations demonstrated the most effective prediction skills, followed by the historical rainfall at each station. Even though only three stations (ST-465008, ST-465010, and ST-500202) presented that historical rainfall was an effective predictor in SLR, it indicated a better performance used with other predictors in MLR and increased with PR. The NINO Index for measuring temperature changes in the Central Pacific Ocean also impacted rainfall in the Phetchaburi Basin. Using SLR, NINO1+2 and NINO3 were effective predictors for two and one station, respectively. However, NINO1+2, when combined with other predictors, exhibited better rainfall forecasting for five stations using MLR and PR, similar to NINO 3. In addition, using NINO3.4 and ONI alone in SLR obtained less reliable forecasting results at all stations than the observations. Significant improvements can be seen with the PR method at five and six stations.

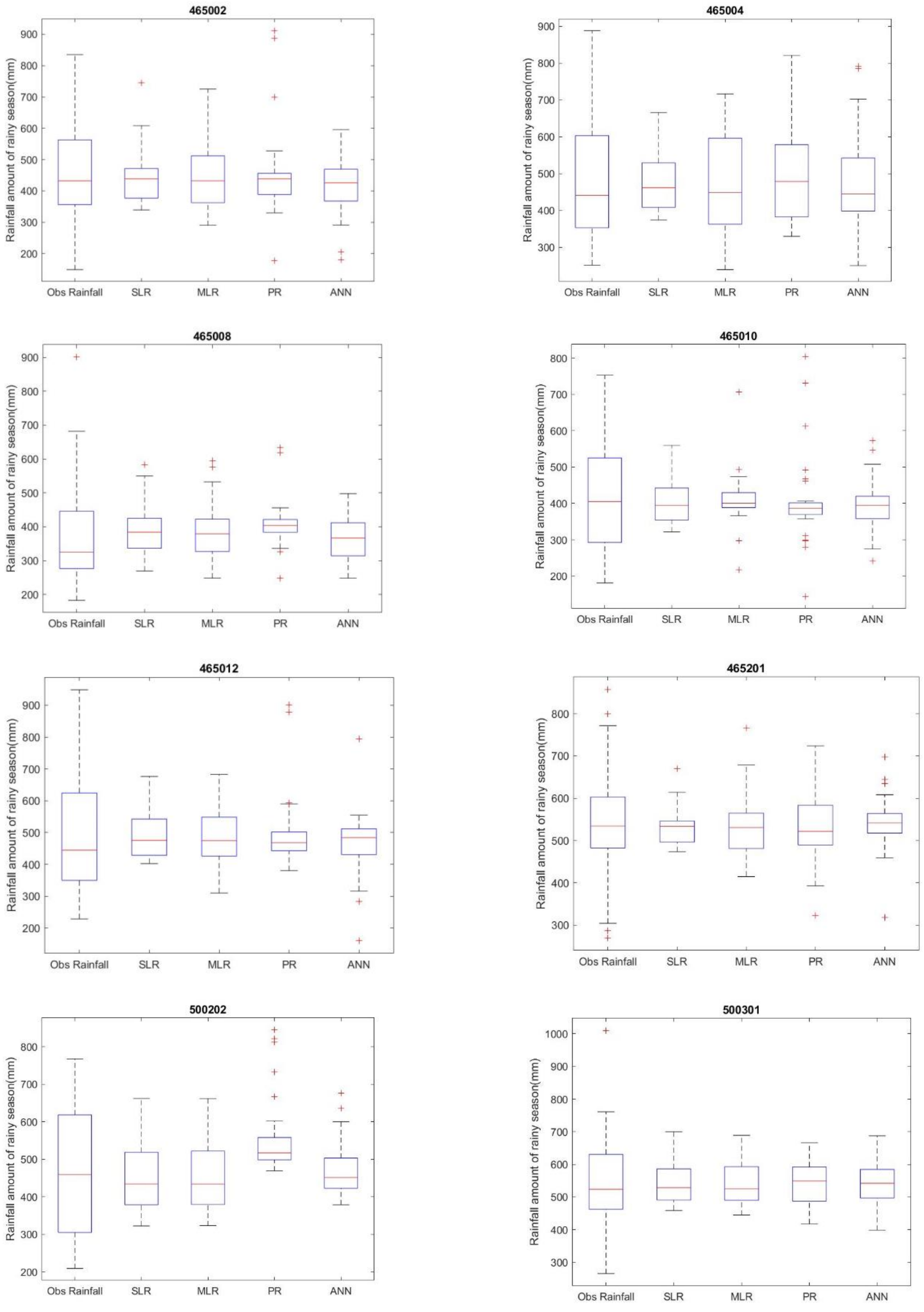

Figure 5. Statistical characteristics of rainfall forecasts and observation

The model performance of all four statistical methods showed that the ANN presented significantly better results than the others at three stations (ST-465010, ST-465012, and ST-500301). The most significant performance of the ANN could be seen at ST-465012 with R and RMSE of 0.59 and 150.4 mm, respectively, whereas the results of the other three models ranged from 0.27–0.37 and 176.9–179.0, respectively. A better regression performance than the ANN was indicated at only two stations, namely ST-465002 and ST-465004. Slightly insignificant differences between MLR and PR were exhibited at ST-465002, with R ranging from 0.52–0.53 and RMSE ranging from 147.5–153.4 mm, whereas rainfall forecasts for ST-465004 using MLR gave outstanding results with the most significant number of predictors (six indices). The statistical characteristics for rainfall during the rainy season at all stations are presented in **Figure 5**. The boxplots show that measures of central tendency between observed and forecasted rainfall in the rainy season at most stations were notably close, except ST-465008 and ST-500202, which showed significant differences with the PR method. Outliers of the forecasted results were also generated when the PR method was applied.

4. Discussion

Rainfall variability in recent years has led to effective water resource management difficulties. The ability to forecast rainfall in the long term at any time step has become vitally important for decision-makers. In this study, various statistical models have been employed along with large-scale climate indices to develop suitable methods for forecasting rainfall in the rainy season with a lead time of 12 to 18 months in advance. Robust evidence supports that DMI is the most effective predictor of rainy season rainfall in the area under study and the selected predictor for all three regressions at most stations. The impact of DMI on the Indochina Peninsula has also been investigated by [22], who revealed that Thailand is one of many countries in their study area to be DMI-sensitive. A linkage between Thailand and DMI was also found in the research work of [23, 24, 25]. Applying DMI with the historical rainfall at each station and the NINO variable significantly increased the ability of linear forecasting models compared to using DMI only. Additionally, the selective predictors applicable to each station could be distinct. Statistical models focus solely on the predictor and predictand data without considering the links between physical attributes. As a result, the selection of predictors for each station or time period may vary due to the high sensitivity of data variance, as seen in [2, 9, 21].

The selection of statistical models in rainfall forecasting is strongly related to large-scale circulation variables. In this study, MEI V.2 and ONI were correlated to local rainfall with non-linearity with PR application. Therefore, large-scale atmospheric indices should be carefully considered for use as predictors in long-lead time rainfall forecasting. The ANN's ability to capture non-linear influences in relationships between large-scale predictors and local rainfall has been widely mentioned in various research works [26,27] and is also presented in this study. However, the forecasting skills exhibited by MLR, PR, and the ANN were insignificantly different at some stations. Therefore, carefully selecting the model and variables to seek the most appropriate methods and predictors is strongly recommended before finalizing the forecasting results. In addition, bias is one of the issues of concern when using statistical models. Statistical models strongly rely on empirical equations to generate forecasting results. Bias can be avoided by removing the dataset used to construct the model before calculation.

5. Conclusion

The objective of this study is to forecast long-term seasonal rainfalls in the Phetchaburi River Basin 12–18 months in advance, using statistical methods with seven atmospheric circulation indices: ONI, DMI, MEI V. 2, NINO4, NINO3.4, NINO3, and NINO1+2, together with historical rainfall. Three statistical methods were used: simple linear regression analysis, multiple linear regression analysis, and polynomial linear regression in combination with k-cross-validation to prevent forecasting bias. In addition, the most suitable period for forecasting was analyzed using the moving window average approach from January to July of the preceding year (12–18 months lead time) to seek the best periods for rainfall forecasting.

In the study area of the Phetchaburi Basin, the Indian Ocean (DMI), and the Pacific Ocean (NINO), sea surface temperature indices were revealed to be the most crucial for forecasting rainfall. The MEI V.2 and ONI, which also reflect climate variability in the Pacific Ocean region, were positively correlated with local rainfall only using non-linear regression. At most stations, non-linear regression indicated better prediction ability than linear regression. However, the initial selection of suitable variables and statistical models is strongly recommended before undertaking long-lead time rainfall forecasting due to the complexity of the relationship between local rainfall and large-scale circulation indices. In addition, the specific selection of predictor periods for rainfall forecasting should be avoided due to the variability involved, as the results show.

6. Acknowledgements

The authors thank the Thai Meteorological Department for providing valuable data.

Author Contributions: Conceptualization, K.S.; methodology, K.S. and N.R.; validation, N.R. and R.T; formal analysis, K.S., N.R. and R.T.; writing—original draft preparation, N.R. and R.T.; writing—review and editing, K.S.; visualization, K.S.; supervision, K.S.

Funding: This research was partially supported by the Faculty of Engineering at Kamphaengsaen Campus, Kasetsart University.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Yumagulova, L.; Vertinsky, I. Climate Change Adaptation and Flood Management in Metro Vancouver Regional Area: Can an Exercise in Herding Cats be successful. *J. Sustain. Dev. Energy Water Environ. Syst.* **2017**, *5*(3), 273-288.

[2] Hossain, L.; Rasel, H. M.; Lmteaz, M. A.; Mekanik, F. Long-Term Seasonal Rainfall Forecasting using Linear and Non-Linear Modelling Approaches: A Case Study for Western Australia. *Meteoro. Atmos. Phys.* **2020**, *132*, 131-141.

[3] Singh, S.; Xiaosheng, Q. Study of Rainfall Variabilities in Southeast Asia using Long-Term Gridded Rainfall and Its Substantiation through Global Climate Indices. *J. Hydrol.* **2020**, *585*, 124320.

[4] Chen, L.; Dool, H.; Becker, E.; Zhang, Q. ENSO Precipitation and Temperature Forecasts in the North American Multimodel Ensemble: Composite Analysis and Validation. *Am. Meteorol. Soc.* **2017**, 1103-1125 .

[5] Adhikari, S.; Liyanaarachchi, S.; Chandimala, J.; Nawarathna, B. K.; Bandara, R.; Yahiya, Z.; Zubair, L. Rainfall Prediction based on the Relationship between Rainfall and El Nino Southern Oscillation (ENSO), *J. Natl. Sci. Found. Sri.* **2010**, *38*(4), 249-255.

[6] Hossain, L.; Rasel, H. M.; Lmteaz, M.A.; Mekanik, F.; Long-Term Seasonal Rainfall Forecasting: Efficiency of Linear Modelling Technique. *Environ. Earth Sci*. **2018**, *77*(280), 1-10.

[7] Jung, J.; Kim, H. S. Predicting Temperature and Precipitation during the Flood Season based on Teleconnection. *Geosci. Lett*. **2022**, *9*(4), 1-37.

[8] De Silva, M.; Hornberge,r G. M. Identifying El Nino-Southern Oscillation Influences of Rainfall with Classification Models: Implications for Water Resource Management of Sri Lanka. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 1905-1929.

[9] Khastagir, A.; Hossain, I.; Anwar, A. H. M. F. Efficacy of Linear Multiple Regression and Artificial Neural Network for Long-Term Rainfall Forecasting in Western Australia. *Meteorol. Atmos. Phys*. **2022**, *134*, 69.

[10] Pontoh, R. S.; Toharudin, T.; Ruchjana, B. N.; Sijabat, N.; Puspita, M. D. Bandung Rainfall Forecast and Its Relationship with Nino 3.4 using Nonlinear Autoregressive Exogenous Neural Network. *Atmosphere*. **2022**, *13*, 302.

[11] Kim, C. G.; Lee, J.; Lee, J. E.; Kim, N. W.; Kim, H. Monthly Precipitation Forecasting in the Han River Basin, South Korea, using Large Scale Teleconnection and Multiple Regression Models. *Water*. **2020**, 12, 1590.

[12] Gnanasankaran, N.; Ramaraj, E. A. Multiple Linear Regression Model to Predict Rainfall using Indian Meteorological Data. *Int. J. Adv. Sci. Technol*. **2020**, *29*(8), 746-758.

[13] Sittichok, K. Seasonal Rainfall Forecasting in Tropical Region Using Statistical Models and Sea Surface Temperatures. *Science and Technology Journal*. **2016**, *5*(3), 33-50.

[14] Abbot, J.; Marohasy, J. Forecasting of Medium-Term Rainfall using Artificial Neural Networks: Case Studies from Eastern Australia. *Engineering and Mathematical Topics in rainfall Intech*. **2017**, https://doi.org/10.5772/intechopen.72619.

[15] Lee, J.; Kim, C. G.; Lee, J. E.; Kim, N. W.; Kim, H. Medium-Term Rainfall Forecasts using Artificial Neural Networks with Monte-Carlo Cross-Validation and Aggregation for the Han River Basin, Korea. *Water*. **2020**, *12*, 1743.

[16] Afshin, S.; Fahmi, H.; Alizadeh, A.; Sedghi, H.; Kaveh, F. Long Term Rainfall Forecasting by Integrated Artificial Neural Network-Fuzzy Logic-Wavelet Model in Karoon Basin. *J. Sci. Res. Essay*. **2011**, *6*, 1200-1208.

[17] Mekanik, F.; Imteaz, M. A.; Gato-Trinidad, S.; Elmahdi, A. Multiple Regression and Artificial Neural Network for Long-Term Rainfall Forecasting using Large Scale Climate Modes. *J. Hydrol*. **2013**, *503*, 11-21.

[18] Acharya, R.; Pal, J., Das, D.; Chaudhuri, S. Long-Range Forecast of Indian Summer Monsoon Rainfall using an Artificial Neural Network Model. *Meteorol. Appl.* **2017**, *26*, 347-361.

[19] Darji, M. P.; Dabhi, V. K.; Prajapati, H. B. Rainfall Forecasting using Neural Network: a Survey. Proceeding of International Conference on Advances in Computer Engineering and Applications (ICACEA), IMS Engineering College, Ghaziabad, India. **2015**.

[20] Liu, Q.; Zou, Y.; Liu, X.; Linge, N. A Survey on Rainfall Forecasting using Artificial Neural Network, Internat. *J. Embed. Syst.* **2019**, *11*(2), 240-249.

[21] Sigaroodi, S.K.; Chen, Q.; Ebrahimi, S.; Nazari, A.; Choobin, B. Long-Term Precipitation forecast for Drought Relief using Atmospheric Circulation Factors: A Study on the Maharloo Basin in Iran. *Hydrol.Earth Syst. Sci*. **2014**, https://doi.org/10.5194/hess-18-1995-2014.

[22] Gao, Q. G.; Sombutmounvong, V.; Xiong, L.; Lee, J. H.; Kim, J. S. Analysis of Drought-Sensitive Areas and Evolution Patterns through Statistical Simulations of the Indian Ocean Diploe mode. *Water*. **2019**, 11, 1302 .

[23] Muangsong, C.; Cai B.; Pumijumnong N.; Hu C.; Cheng H. An Annual Laminated Stalagmite Record of the Changes in Thailand Monsoon Rainfall over the Past 387 Years and Its Relationship to IOD and ENSO. Quat. Int. **2014**, *349*, 91-97.

[24] Ha, K. J.; Seo, Y. W.; Lee, J. Y.; Kripalani, R. H., Yun, K. S., Linkages between the South and East Asians Summer Monsoons: A Review and Revisit. *Clim. Dyn*. **2017**. DOI 10.1007/s00382-017-3773-z.

[25] Hoell, A.; Harrison, L.; Indian Ocean Dipole and Precipitation. *Agroclimatology Fact Sheet Series (Famine Early Warning Systems Network)*. **2021**, *3*, 1-2.

[26] Badr, H. S.; Zaitchik, B. F.; Guikema, S. Application of Statistical Models to the Prediction of Seasonal Rainfall Anomalies over the Sahel. *J. Appl. Meteorol. Climatol.* **2014,** *53*, 614-636.

[27] Golian, S.; Murphy, C.; Wilby, R. L.; Matthews, T.; Donegan, S.; Quinn, D. F.; Harrigan, S. Dynamical – Statistical Seasonal Forecasts of Winter and Summer Precipitation for the Island of Ireland. *Int. J. Climatol.* **2022**, https://doi.org/10.1002/joc.7557.