Spatial interpolation methods for estimating monthly rainfall distribution in Thailand

N. Chutsagulprom[a,d,e], K. Chaisee[c,d,e,*], B. Wongsaijai[a,d,e], P. Inkeaw[b,d,e], C. Oonariya[f]

[a]*Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand, 50200*

[b]*Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand, 50200*

[c]*Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand, 50200*

[d]*Advanced Research Center for Computational Simulation, Chiang Mai, Thailand, 50200*

[e]*Centre of Excellence in Mathematics, CHE, Si Ayutthaya Rd., Bangkok, Thailand, 10400*

[f]*Climate Center, Thai Meteorological Department, Sukhumvit Rd., Bangkok, Thailand, 10260*

*Corresponding author

Email address: kuntalee.chaisee@cmu.ac.th (K. Chaisee)

Abstract

Spatial interpolation methods usually differ in their underlying mathematical concepts, each with inherent advantages and drawbacks depending on the properties of data. This paper, therefore, aims to compare and evaluate the performances of well-established interpolation techniques for estimating monthly rainfall data in Thailand. The selected methods include the inverse distance-based method, multiple linear regression (MLR), artificial neural networks (ANN), and ordinary kriging (OK). The technique of searching nearest stations is additionally imposed for some aforementioned schemes. The *k*-fold cross-validation method is exploited to assess the efficiency of each method, then the metric scores, RMSE, and MAE are used for comparisons. The results suggest the ANN might be the least favorite as it underperforms in many folds. While the OK method provides the most accurate prediction, inverse distance weighting (IDW), particularly inverse exponential weighting (IEW), and MLR are considerably comparative. Overall, IEW is plausible for the monthly rainfall estimation of Thailand because it is less computationally expensive than the OK and is flexible computation.

Keywords: artificial neural networks, inverse distance weighting, multiple linear regression, ordinary kriging, rainfall data, spatial interpolation

Introduction

Motivation

One of the main sectors in Thailand's development is agriculture, based on geographic location and the availability of natural resources. Rainfall data are inevitably vital information for both hydrological and agricultural management. The spatial analysis of this resource is undoubtedly essential for the implementation of the national agricultural strategic plan. In general, rainfall data are recorded thoroughly at rain gauge stations located sparsely across the region of interest. Nevertheless, due to many factors such as malfunctioning equipment or severe environmental conditions, missing values during the data collection process are typical phenomena. According to the Thai Meteorological Department, there are 75 rain gauge stations in total located in the area of 513,120 square kilometers that may be insufficient to provide the necessary information. It is, therefore, necessary to apply spatial interpolation techniques so that precipitation patterns can be fully captured. This research intends to determine an optimal interpolating method for monthly rainfall estimation in Thailand using some neighboring rainfall data and the location of rain gauge stations as predictors. We impose cross-validation to evaluate the performance of the methods. The best-suited method for monthly rainfall estimation in Thailand is then obtained from the comparison.

Related work

There have been various spatial interpolation methods in which the underlying premise behind most techniques are by assigning weights to each of the available data points. Inverse distance weighting (IDW) is one of the deterministic interpolation methods. It is a relatively straightforward implementation where the weights are inversely proportional to the power of

the distance between the locations. Chen & Liu (2012) applied the IDW method to estimate the rainfall data in the middle of Taiwan and studied the relationship between the interpolation accuracy and two related factors, including the power value and a radius of influence. Successful applications of IDW and its modifications are presented in different places, such as the mainland of China (Chen *et al*. 2010), the eastern Mediterranean (Kurtzman *et al.* 2009), and Texas, US (Kong & Tong 2008). Besides IDW, weighting can be assigned in several ways. The inverse exponential weighting (IEW) method is another inverse distance-based method where a negative exponential function replaces the inverse o distance in the traditional IDW. The IEW approach is commonly used in the field of quantitative geography for surface generation (Goodchild & Janelle 2004). By contrast, kriging interpolation is a probabilistic-based technique providing the estimation of unknown quantities in terms of a linear combination of nearby sampled points with the kriging weights being evaluated through the semivariogram. Ordinary kriging (OK) is one of the most common kriging approaches in practice. It is applied extensively to various geoscience variables, including temperature, solar radiation, and precipitation. Particularly, it is usually chosen to analyze spatial patterns of rainfall intensity across a region of interest. Murthy & Abbaiah (2007) generated a map of spatial variability of rainfall in India employing OK, where the co-kriging is further used to improve the prediction, with the elevation being included as a correlated variable. Suhaila & Jemain (2012) applied OK with the exponential semivariogram to identify locations of high intensity of daily rainfall over Peninsular Malaysia. Multiple linear regression (MLR) is statistical modeling used to investigate the relationship between several explanatory variables and a response variable. Although it is generally exploited as a predictive analysis means by fitting a linear function to the data, it is utilized as an interpolation tool as published in (Bostan *et al.* 2012; Sharifi *et al.* 2019). An artificial neural network (ANN) composes of the structures that imitate the functioning of a human brain. Basic elements of the ANN are neurons that are interconnected by weighted links to form a network. Each neuron in the network performs

some function on the summed value of the neighboring data. An example of the ANN with application to rainfall data includes Teegavarapu & Chandramouil (2005). They used the ANN to estimate missing precipitation values in Kentucky, USA, by treating the latitude, longitude, and elevation values at each station as the input. Amiri & Mesgari (2017) applied the ANN to estimate annual rainfall in Northwest Iran. Several studies focus on searching for the optimal approach for spatial precipitation interpolation. Tabios *et al.* (1985) utilized the Thiessen polygon, IDW, and OK to interpolate annual rainfall data in the North Central continental United States based on a 30-year dataset from 29 stations. It appeared that the OK is superior to other chosen techniques. Abtew *et al.* (1993) employed various spatial interpolation methods to estimate the rainfall where the OK outperformed other interpolation schemes. On the contrary, Yang *et al.* (2015) compared and assessed four spatial interpolation techniques, namely IDW, ANUDEM, Spline, and kriging, in the case of daily rainfall data over the greater Sydney region. In accordance with their error analysis, it is suggested that IDW has the best predictive capability relative to other methods, while kriging produced slightly different seasonal patterns. Using monthly precipitation records from 20 sites situated in central Chile, Barrios *et al.* (2018) compared five different interpolation schemes under the consideration of the radius of influence. The overall best performance can be obtained from the ANN, MLR, and the modified version of the IDW method.

In this paper, our purpose is to provide and compare diverse types of interpolating techniques for estimating the monthly rainfall of Thailand based on historical rainfall data. We focus on well-known and practical interpolation methods; IDW, IEW, MLR, ANN, and OK. The comparison of root mean square and mean absolute error in k-fold cross-validation are used to evaluate the efficiency of the methods.

Spatial Interpolation Methods

As mentioned earlier, the estimate at the non-visited site of most interpolation schemes can primarily be expressed in terms of a combination of observed data and only differ in techniques used to find assigning weights. Let us assume that $\{Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2), \ldots, Z(\boldsymbol{x}_n)\}$ is a set of samples at sites $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. The estimate Z^* at the unmeasured location $\boldsymbol{x}$ is therefore given by

$$Z^*(\boldsymbol{x}) = \sum_{i=1}^{n} w_i Z^*(\boldsymbol{x}_i) \tag{1}$$

with an imposed condition

$$\sum_{i=1}^{n} w_i = 1, \tag{2}$$

where w_i's are the normalized weights that must be determined.

Inverse distance-based method

IDW is one of the commonly used techniques that follow the above concept. It is a multivariate interpolation approach where its comprehensible assumption lies in the fact that weight coefficients diminish as the distance increases. The weight values in Eq. (2) can be computed from the following equation

$$w_i = \frac{d_i^{-p}}{\sum_{i=1}^{n} d_i^{-p}},$$

where p can be chosen arbitrarily, and $d_i = d(\boldsymbol{x}, \boldsymbol{x}_i)$ is the distance between $\boldsymbol{x}$ and $\boldsymbol{x}_i$. The main factor affecting the accuracy of this method is the value of the power parameter p. The most common choice is, however, when p is 2, a so-called inverse distance squared method. As p increases, the quantity at the unmeasured site is contributed by nearer observation points. This is somewhat similar to the estimation via the polygon method.

The success of the inverse distance-based method that has been broadly exploited in literature is the inverse exponential weighting method (IEW), which is defined by the mean of a negative exponential function

$$w_i = \frac{\exp(-pd_i)}{\sum_{i=1}^{n} \exp(-pd_i)}.$$

The IEW is similar to the IDW; therefore, the most commonly used value for p is also 2, which is typically tested before arriving at a satisfactory value.

Multiple linear regression

MLR is one of the widely used methods in forecasting. It attempts to predict a response (dependent) variable from several explanatory (independent) variables. In other words, it explains the response variable's behavior based on the linear association between the variables. Consider explanatory variables $x_i, i = 1,2,...,r$ and response variable y; the multiple linear regression model can be formulated as follows:

$$y = \beta_0 + \sum_{i=1}^{r} \beta_i x_i + \varepsilon.$$

In data analysis with n observations, the MLR model can be expressed in the form of vector and matrix as

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}$ is $(n + 1) \times 1$ vector of regression coefficient parameters, and $\boldsymbol{\varepsilon}$ is $n \times 1$ vector of random errors. The noise term $\boldsymbol{\varepsilon}$ is assumed to be drawn from the normal distribution such that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$, leading to $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 \mathbf{I})$. Details of MLR can be found in (Bostan *et al.* 2012; Sharifi *et al.* 2019).

Artificial neural networks

an ANN is a computational model inspired structurally and functionally in biological neural networks (Coulibaly & Evora 2007). Similar to the human brain, the ANN is a web of neuron nodes. As shown in Figure 1, a neuron is a processing unit that receives inputs and subsequently manipulates these inputs in order to generate outputs. Given the inputs x_i, $i = 1, 2, ..., n$, a linear combination of the input and its associated weight is initially generated and later fed to an activation function $\phi(\cdot)$ to create an output y, that is

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i\right) \tag{3}$$

where w_i are parameters of the model. A well-known structure of ANN is the multilayer perceptron in which the neurons are arranged in a layer, with the output of one layer serving as the input to the next layer and possibly other layers. The model parameters w_i can be optimized by using the Levenberg-Marquardt training algorithm.

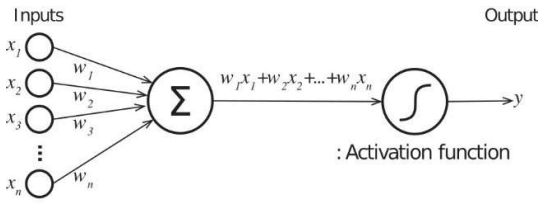


Figure 1: Schematic illustration of a neuron.

Ordinary kriging

OK is a geostatistical interpolation technique associated with the best linear unbiased estimator as it involves the minimization of error of the estimation (Cressie 1990). The spatially distributed data are considered samples of a random field, which allows some probabilistic properties applicable to estimate variables at unvisited locations. The estimate can be calculated in the manner of Eq. (1) along with the constraint in Eq. (2), while the weights can be found under the intrinsically stationary assumption. The variance of the estimation can accordingly

be determined in terms of the semivariogram that represents the spatial autocorrelation structure of the observations (Cressie 1988), that is

$$Var[Z^*(\boldsymbol{x}) - Z(\boldsymbol{x})] = -\sum_{j=1}^{n}\sum_{i=1}^{n} w_j w_i \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) + 2\sum_{i=1}^{n} w_i \gamma(\boldsymbol{x}_i - \boldsymbol{x}). \quad (4)$$

By applying the method of Lagrange multiplier to the objective function Eq. (4) subject to Eq. (2) implies

$$\sum_{j=1}^{n} w_j \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) + \mu = \gamma(\boldsymbol{x}_i - \boldsymbol{x}),$$

where μ is the Lagrange multiplier. Therefore, unlike the IDW, the determination of kriging weights relies strongly on the semivariogram. The empirical semivariogram can be constructed according to (Matheron 1963)

$$\hat{\gamma}(\mathbf{d}) = [1/n(\mathbf{d})] \sum_{i=1}^{n(\mathbf{d})} (Z(\mathbf{x}_i) - Z(\mathbf{x}_i + \mathbf{d}))^2, \quad (5)$$

where $n(\mathbf{d})$ is the number of pairs of data locations at a separation vector $\mathbf{d}$. Under the isotropic assumption, the semivariogram in Eq. (5) can be written as $\hat{\gamma}(d)$ when d is the magnitude of a vector $\mathbf{d}$. Some continuous function is thereafter imposed in order to fit this empirical semivariogram. Undoubtedly care must be taken when specifying the fitted semivariogram function. Here various parametric semivariograms are opted for quantifying the spatial autocorrelation of our dataset, but we only present the exponential model as its overall performance is superior to others. The model is expressed as

$$\gamma(d, \mathbf{c}) = \begin{cases} 0, & d = 0, \\ c_0 + c_1\left(1 - \exp\left(-\dfrac{3d}{c_2}\right)\right) & d \neq 0, \end{cases}$$

where c_0, c_1 and c_2 are nonnegative parameters to be chosen. The least square method is subsequently applied to gain an optimal set of parameters.

Results and Comparisons

Study area and data

Thailand (15.8700◦N, 100.9925◦E) is a country in Southeast Asia and has a tropical climate, thereby the country is becoming generally hot in particular between March-May. The monsoon season runs from mid-May to October. Besides, the monsoon causes mild weather and abundant rain in the southern region, where the topography is the peninsula between the Andaman Sea and the South China Sea. Thailand has 75 rain gauge stations situated across the country. Due to missing data in some stations, the data used in this study is from 67 gauging stations. Throughout the study, we use the historical rainfall data from 1983 to 2018 (36 years) from these 67 stations, as illustrated in Figure 2. Figure 3 presents characteristics of monthly and annual residuals that are also known as monthly and annual rainfall anomalies.

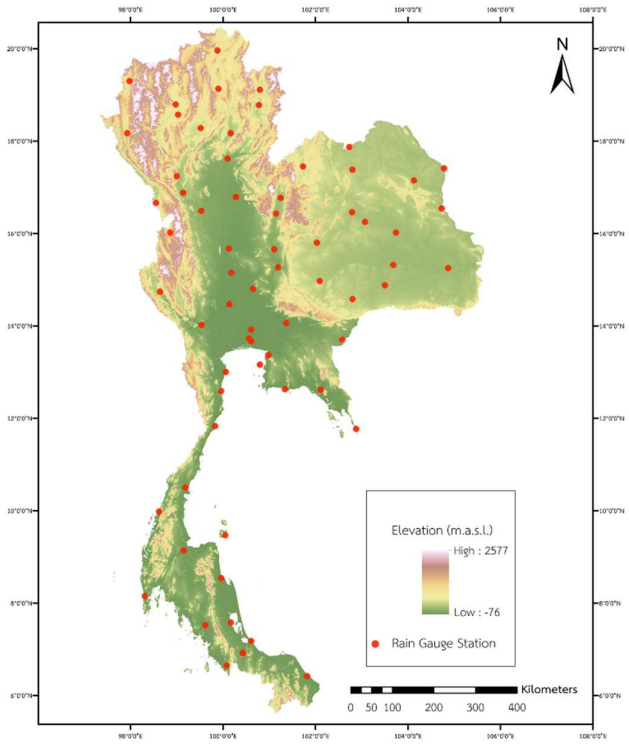


Figure 2: Locations of the rain gauge stations of Thailand.

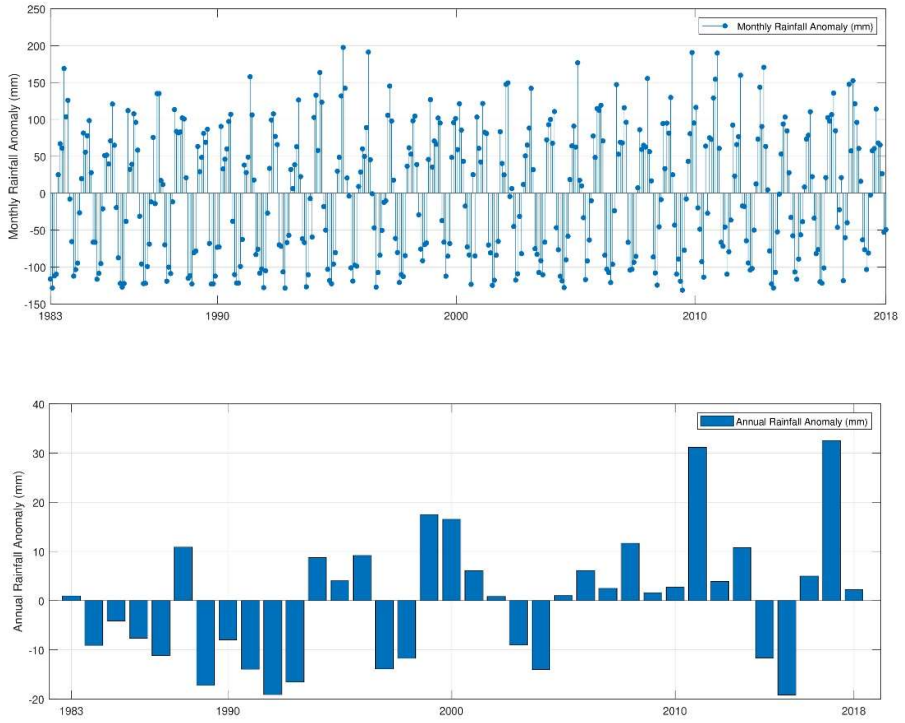


Figure 3: Monthly rainfall anomaly (Top) Annual rainfall anomaly (Bottom) in Thailand during 1983-2018.

The search of nearest stations

Under some situations, the estimates can be inaccurate when the distance between stations is not physically associated due to sampling errors. We can reduce such errors by considering nearby stations that are in the vicinity of our target station. The estimates at unsampled locations derived from the IDW, MLR, and ANN methods are evaluated by taking the specific number of closest neighboring stations (Barrios *et al.* 2018) into account, as can be seen in Figure 4. As a result, it provides a minor modification of Eq. (1) in which, instead of using *N,* the number of all stations, we now replace it with the specific number of neighboring sites. In this study, we separate into two cases depending on the number of nearby stations used to compute the estimates at the unvisited locations, 3 and 9 stations. In contrast to the OK method, at least 50 neighboring data should be acquired in order to gain a scrupulous geostatistical analysis, especially for the construction of an appropriate semivariogram (Webster & Oliver

1993). Henceforth, all measured data are exploited to estimate rainfall concentration in the ordinary kriging, while the search nearby stations are applied to the others.

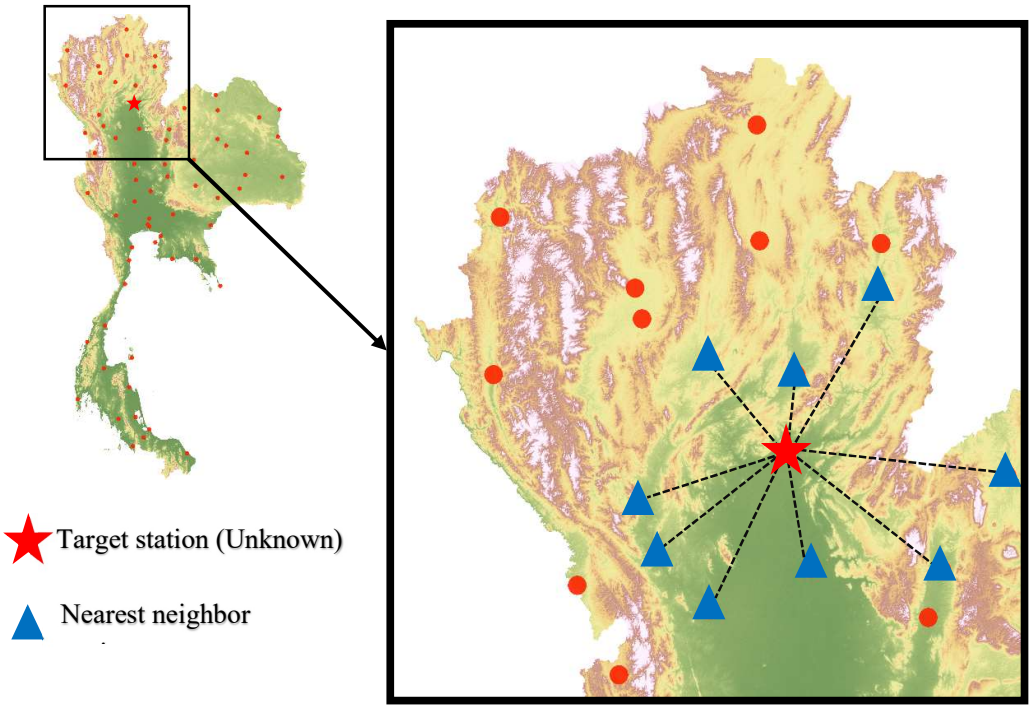


Figure 4: The Search of the nearest stations.

Cross-validation and evaluation criteria

The *k*-fold cross-validation technique is applied to evaluate the quality of five distinct interpolation methods. The principal concept of such a method is to randomly but evenly partitioning the dataset into *k* sub-collections. Each time, one of these sub-collections is used as a validation dataset, while the remaining data are treated as a training dataset. In this study, we use *k* being 10. In terms of the metric used to assess the method performance, the root mean square error (RMSE) and the mean absolute error (MAE) are employed. The RMSE is a quadratic scoring rule which is given by the following equation

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z^{*}(x_i) - Z(x_i))^2}, \qquad (6)$$

where n is the sample size of the testing dataset, $Z^{*}(x_i)$ and $Z(x_i)$ are interpolated and observed

values, respectively. While the MAE is a linear scoring that measures the average magnitude of the errors without considering their direction. The MAE can be formulated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Z^{*}(x_i) - Z(x_i)\right|, \tag{7}$$

In each fold, we obtain the errors of monthly rainfall in each station. Therefore, in each month, we calculate the RMSEs and MAEs expressed in Eqs. (6) and (7). The interpolations are run over the rainfall data is over 36 years, so that the errors of monthly averages are obtained from a total of 36 years. To represent the performance of the methods, we then use the average of the RMSEs and the MAEs obtained from all folds.

Predictive ability of interpolation

Inverse distance-based method

We first start by comparing the weighted distance interpolations, IDW and IEW. The range of p values of both methods is $\{1, 2, 3, 4\}$. Table 1 displays the errors obtained from these two techniques. It can be observed that when the number of nearby stations used in the estimation is imposed, relatively low values of p yield a more accurate estimation in both approaches. Specifically, the lowest values of RMSE and MAE are achieved when $p = 2$ with 9 neighboring stations, with RMSE being 76.57 for IDW and 76.14 for IEW. Although the best performance in the case when all testing data is used is still gained when $p = 2$, the results acquired from $p = 1$ are unsatisfactory.

Table 1: Predictive evaluation of IDW and IEW.

Model	p	3 neighboring		9 neighboring		All stations	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
IDW	1	79.11	57.82	76.29	56.06	82.85	62.83
	2	79.94	58.16	76.57	55.92	77.42	57.12
	3	81.53	59.00	78.50	56.99	78.48	57.07
	4	83.18	59.09	80.79	58.25	80.71	58.21
IEW	1	78.67	57.71	76.23	56.23	77.53	57.59
	2	78.84	57.56	76.14	55.79	76.24	55.93
	3	80.03	58.07	77.61	56.38	77.51	56.30
	4	81.59	58.84	79.65	57.43	79.58	57.38

MLR interpolation

We use the monthly rainfall data in each rain gauge station as the response variable with two different models regarding exploratory variables. The first model, named the Lat-Long model, uses the coordinate (latitude and longitude) of the observed stations as exploratory variables. On the other hand, the second model, referred to as the Nearest Neighbor model, receives the observed values in the k neighboring stations (i.e., $k \in \{3, 9\}$) as exploratory variables. It has also been used in (Teegavarapu & Chandramouli 2005) for estimating monthly precipitation. The results are shown in Tables 2 - 3. Similar to the IDW and IEW, the MLR approach is significantly affected by the number of neighboring stations.

Table 2: Predictive evaluation of MLR.

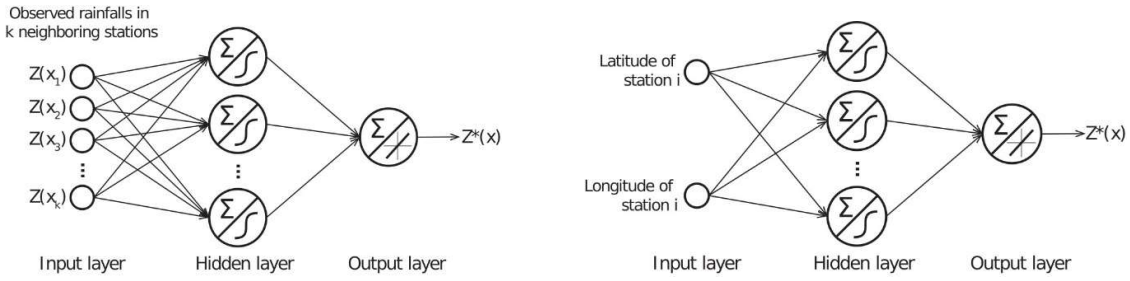
Model	3 neighboring		9 neighboring	
	RMSE	MAE	RMSE	MAE
Lat-Long	131.03	91.68	77.46	57.90
Nearest	84.01	60.93	107.25	75.66

ANN interpolation

The Lat-Long and Nearest Neighbor models previously mentioned in the MLR method are also adopted in ANN, where the exploratory variables can be interpreted as input. Except for the Lat- Long model, instead of using the coordinate of all stations as input, the latitude and longitude of the target station are served as input. In the ANN approach, these two models consist of one hidden layer and produce the estimated precipitation for the target station as the output. In the hidden layer, the activation function employed is the familiar sigmoid function. The number of hidden nodes is determined by selecting an optimal number in set {2, 4, 6, 8, 10}. For the output layer, a linear function is used as the activation function. Figure 5 shows the results of the Nearest Neighbor and Lat-Long models. The results in Table 2 show that when 3 neighboring stations are used, the Nearest Neighbor model is better than the Lat-Long model. On the contrary, when 9 neighboring stations are employed, the Lat-Long model performs better. Interestingly, in Table 3, the ANN interpolation shows that using 3 neighboring stations is considerably good. The key of this method is the number of hidden nodes, as the results suggest that the Lat-Long model requires more nodes than the Nearest Neighbor model.

Table 3: Predictive evaluation of ANN.

Model	Hidden nodes	3 neighboring		9 neighboring	
		RMSE	MAE	RMSE	MAE
Lat-Long	2	103.82	73.36	184.86	105.57
	4	93.50	67.49	108.13	77.59
	6	91.02	65.76	111.18	80.97
	8	**90.62**	**65.46**	113.71	83.39
	10	91.11	65.89	113.53	83.12
Nearest	2	165.20	100.55	132.23	90.15
	4	114.55	**81.12**	**112.20**	**83.30**
	6	**113.82**	82.48	116.42	87.91
	8	118.73	86.19	125.25	93.30
	10	123.81	89.63	129.61	97.34



(a) Nearest Neighbor architecture (b) Lat-Long architecture

Figure 5: Architectures of the artificial neural networks used for estimating monthly rainfall.

OK

Rainfall data can be inconsistent with the normal distribution, which may cause the semivariogram to exhibit a proportional effect and produce distortion to cover the inherent structure. Thus, normality conversion is performed before applying OK interpolation. We apply the Box-Cox transformation method to provide the normality. The transformation is given by the values λ; hence the data can be expressed as follows.

$$Z^* = \frac{Z^\lambda - 1}{\lambda},$$

where $\lambda \neq 0$. Typically, values of λ are chosen from the set of 0.1, 0.25, 0.5, and 1 as they are well-known transformations, which have been suggested in geostatistical radar-gauge combination (Schuurmans *et al.* 2007). It is known as the log transformation when $\lambda = 0.1$, the square root-square root transformation when $\lambda = 0.25$, the square root transformation when $\lambda = 0.5$, and no transformation when $\lambda = 1$. Nevertheless, Erdin *et al.* (2012) has suggested that the optimal values of λ can be chosen from the interval [0.2, 1]. Unlike the previous methods, the OK interpolation requires all available data to exploit the method. The performance of the OK method with varying values of λ is shown in Table 4. According to the quantitative scorings, the predictive accuracy when λ being 0.25 is the most preferable, with RMSE being 73.45 and MAE being 53.03. It exhibits about 24% improvement when compared to λ being 1. It strongly suggests that data transformation is essential as the errors are relatively large when there is no transformation.

Table 4: Predictive evaluation of OK.

λ	0.10	0.25	0.50	1.00
RMSE	76.29	**73.45**	73.49	96.31
MAE	53.23	**53.03**	53.36	75.73

Comparisons

In this subsection, the efficiency and accuracy of the selected interpolation approaches are compared through the RMSE and MAE scorings. Here the IDW is excluded as it has already been determined that, among the inverse distance-based methods used in this work, the IEW has a slightly superior estimation ability to the IDW. The IEW is, therefore only chosen for this comparison study. To examine the efficiency of the methods, a comparison of RMSE and MAE in each fold is illustrated in Figure 6. The errors in each fold are comparable; however, the

ANN method's errors are relatively high. The errors in fold 7 are relatively low, whereas they are considerably high in fold 8. We also display the variation of the RMSE values within the months from each station, particularly stations in fold 7 and fold 8 in Figures 7 and 8, respectively. One can see the similarity of trend patterns over the months in fold 7, in which the RMSE values are lower in the dry season and are rising in the wet season. In contrast, there is an unusual trend in fold 8, particularly at stations Phatthalung-Agromet, Phatthalung Agromet, and Satun, where the RMSE values are high in November-December. It is because these stations are located in the south, where the wet season is extended, and monthly rainfall is higher than in the other parts of Thailand (see Figure 2). The comparison among the interpolation methods shows no clear and robust evidence to conclude that one particular method outperforms the others. Nevertheless, the ANN is a slightly deficient approach to estimate the monthly rainfall data in Thailand as it shows inferior performance in many cross-validation folds, as shown in Figure 6. This is in contrast with (Teegavarapu & Chandramouli 2005) where the ANN appeared to be the optimal approach for estimating the missing monthly precipitation. Furthermore, Barrios *et al.* (2018) found that ANN and MLR provided the best performance, especially when it comes to an estimation of the variable of interest in areas with complex topography. An explanation of deriving dissimilar conclusions might be due to the size of network density exploited in the methods. In this work, the network density is approximately 0.13 stations per 1000 square kilometers, whereas that of (Teegavarapu & Chandramouli 2005) being about 1.3 stations per 1000 square kilometers. Overall, the OK is considered the most efficient method according to the lowest errors, while the IDW, IEW, and MLR based on 9 neighboring stations are comparatively accurate. It is worth noting that the OK method uses all available data points. It is hence not surprising that it gives the smallest errors. As there are no promising approaches, we suggest the IDW and IEW approach due to their flexibility and computational simplicity.

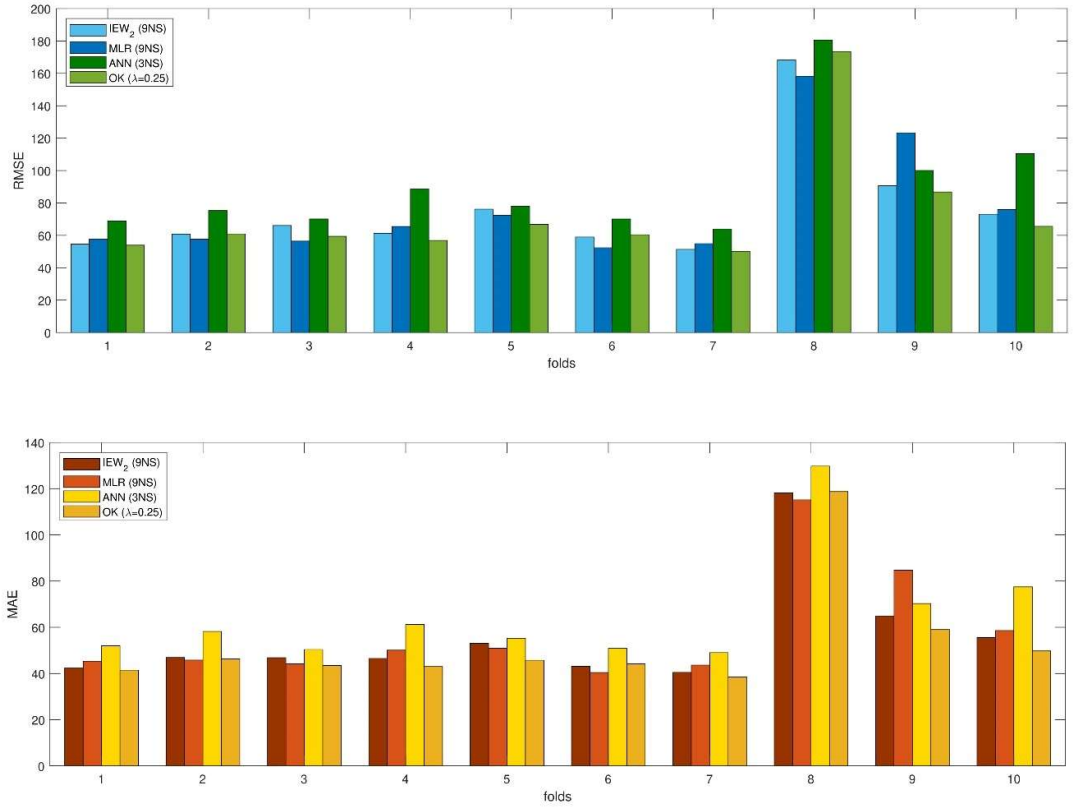


Figure 6: RMSE (Top) and MAE (Bottom).

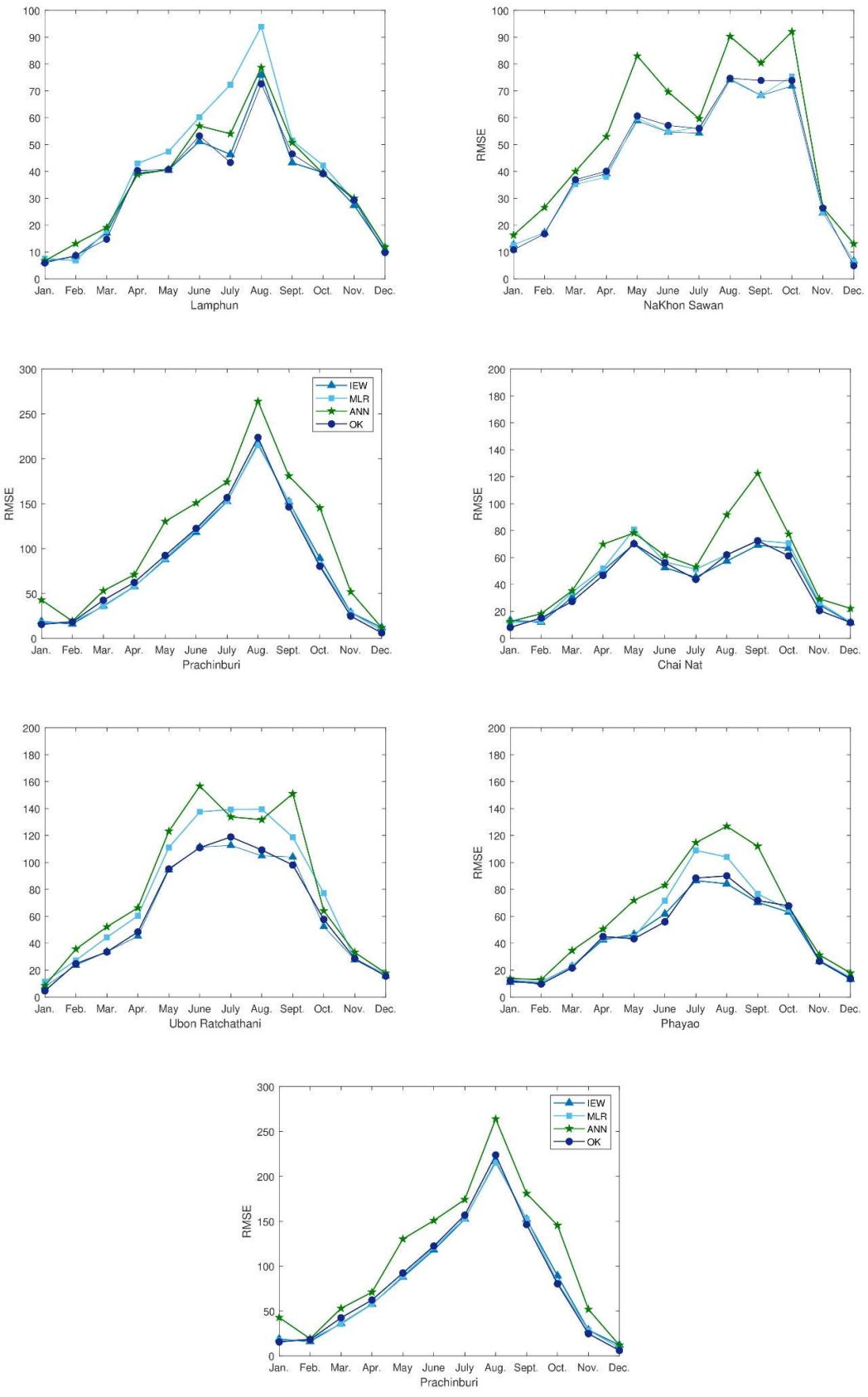


Figure 7: RMSEs of different methods in each moth of 7 rainfall stations in the fold 7.

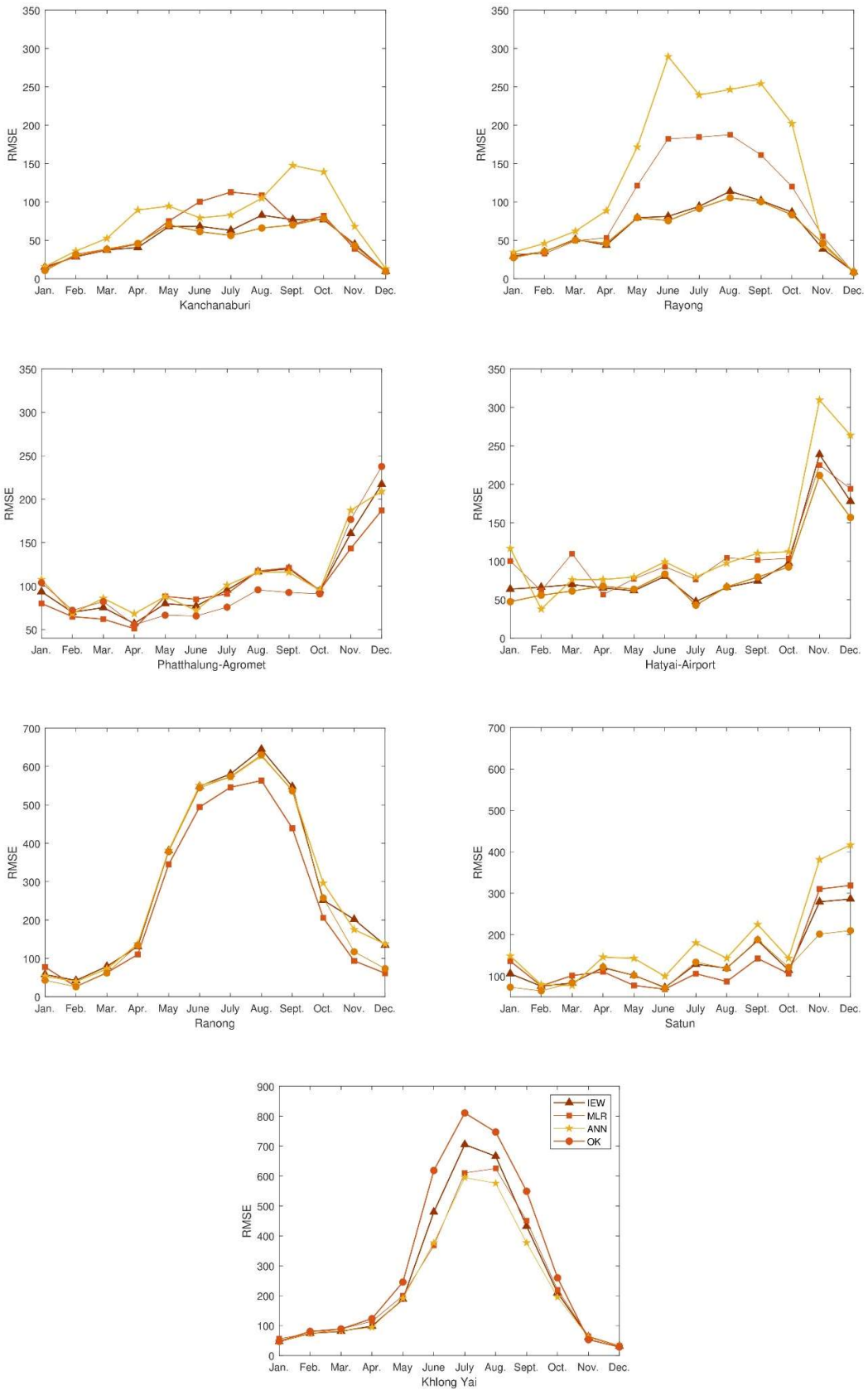


Figure 8: RMSEs of different methods in each month of 7 rainfall stations in the fold 8.

Discussion

Under the characteristic of climate scenarios in different areas over Thailand, it can be classified into 4 categories: (1) the upper portion is a high-land and dry area located in the northern and northeastern parts (2) the central portion is a wet-land area situated in the central part (3) the lower portion is a coastal area located in the southern part, and (4) the most of mountain and mountainous regions are located along with western Thailand and a kind of border from Myanmar. Figures 7-8 show the RMSE variation of the annual cycle of precipitation averaged over the selected point stations around Thailand. The annual cycles averaged over 36 years from 1983 to 2018. The discussion will be arranged as: (1) most of upper Thailand is mountainous and high-land areas located between Latitude 13N and 22N (Figure 7: all stations except Petchaburi station). The high value of the RMSE variation in May and August to September correlated with the impact of monsoon rainfall oscillations, Inter-Tropical Convergence Zone (ITCZ). (2) Lower Thailand is a part of the Indochina Peninsula, the most coastal area, and lies along both sides of the Indian and Pacific Oceans (Figure 8: Phattalung-Agromet, Hatyai-Airport, and Satun stations). It is located between Latitude 5˚N and 12˚N in tropical rain forests. The peak of RMSE variations is in May to September and November to December, reflecting the influence of monsoon rainfall and location along the oceans. All algorithm shows the same pattern of RMSE variation curves. The high value of variations is consistent with the peak of the monsoon season. (3) There are two gauge-based under the total influence of monsoon rainfall season, which is from mid-May to mid-October, including Klong Yai station (102.88˚N, 11.77˚E) and Ranong station (98.62˚N, 9.98˚E). The algorithm in each interpolation shows the normal curve-like of RMSE variation with the high value at the peak of the monsoon season from August to September. (4) The mountainous is one of the topography effects and has a significant influence on the rainfall interpolation. There are two gauge-base stations located near and behind the mountainous area (Phetchburi

(100.06˚N,13˚E) and Kanchanaburi (99.54◦N,14.02◦E) stations). The RMSE variation shows a curve of chaotic pattern in which the peak of fluctuation covers the whole monsoon season, from May to October.

Conclusion

This study uses several interpolation schemes, including the inverse distance-based methods, MLR, ANN, and OK, to seek the optimal monthly rainfall estimation in Thailand. The data used in the computation is the amount of the monthly rainfall collected from the 67 rain gauges in Thailand over 36 years from 1983-2018. The *k*-fold cross-validation is used to assess the performance of different spatial interpolation methods. According to statistical metrics used in the study, the OK generates the most accurate estimation, while the ANN is the least favorable. The IEW and the MLR are not significantly different, but the IEW is more desirable. The OK might be the best choice for high accuracy, but we also recommend the IEW because it has less computational demand than the OK. Besides, it is essential to note that the number of rain gauge stations or the density of rain gauges over the study area can be vital in estimation. Lastly, we might consider taking other factors such as the influence of topography and seasonal monsoon behaviors mentioned earlier into account to improve the estimation accuracy for future work.

Funding This research was supported by Centre of Excellence in Mathematics, Thailand, the Commission on Higher Education, Thailand, the Thailand Research Funds (TRF), Thailand under Grant No. RDG6030003, and Chiang Mai University, Thailand.

Author contributions

N. Chutsagulproma: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, roles/writing of original draft, writing of the review, and editing.

K. Chaiseec: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, roles/writing of original draft, writing of review, and editing.B. Wongsaijai: investigation, writing of the review, software, and editing. P. Inkeaw: investigation, writing of the review, software, and editing. C. Oonariya: investigation, data curation, resources, and editing.

Data availability All data in this study are from the Climate Center, Thai Meteorological Department.

Code availability We wrote all the original programs with MATLAB for the analysis in this work, and we are willing to share them upon request.

Compliance with ethical standards

Competing interests The authors declare that they have no competing interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable

References

Abtew, W., Obeysekera, J. & Shih, G. 1993 Spatial analysis for monthly rainfall in south Florida. *JAWRA Journal of the American Water Resources Association*, **29** (2), 179–188.

Amiri, M. A. & Mesgari, M. S. 2017 Modeling the spatial and temporal variability of precipitation in northwest Iran. *Atmosphere*, **8** (12).

Barrios, A., Trincado, G. & Garreaud, R. 2018 Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile. *Forest Ecosystems*, **5** (1), 28.

Bostan, P. A., Heuvelink, G. B. M. & Akyurek, S. Z. 2012 Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. *International Journal of Applied Earth Observation and Geoinformation*, **19** 115–126.

Chen, D., Ou, T., Gong, L., Xu, C.-Y., Li, W., Ho, C.-H. & Qian, W. 2010 Spatial interpolation of daily precipitation in china: 1951–2005. *Advances in Atmospheric Sciences*, **27** (6) 1221–1232.

Chen, F.-W. & Liu, C.-W. 2012 Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy and Water Environment*, **10** (3), 209–222.

Coulibaly P. & Evora, N. D. 2007 Comparison of neural network methods for infilling missing daily weather records. *Journal of hydrology*, **341** (1-2), 27-41.

Cressie, N. 1988 Spatial prediction and ordinary kriging. *Mathematical geology*, **20** (4), 405-421.

Cressie, N. 1990 The origins of kriging. *Mathematical geology*, **22** (3), 239-252.

Erdin, R., Frei, C. & Künsch, H. R. 2012 Data transformation and uncertainty in geostatistical combination of radar and rain gauges. *Journal of Hydrometeorology*, **13** (4), 1332–1346.

Goodchild, M. F. & Janelle, D. G. 2004 *Spatially integrated social science*. Oxford University Press.

Kong, Y.-F. & Tong, W.-W. 2008 Spatial exploration and interpolation of the surface precipitation data. *Geographical Research*, (5) 15.

Krishna Murthy, B. R. & Abbaiah, G. 2007 Geostatistical analysis for estimation of mean rainfalls in Andhra Pradesh, India. *International Journal of Geology*, **1**, 35–51.

Kurtzman, D., Navon, S. & Morin, E. 2009 Improving interpolation of daily precipitation for hydrologic modelling: spatial patterns of preferred interpolators. *Hydrological Processes: An International Journal*, **23** (23) 3281–3291.

Matheron G. 1963 Principles of geostatistics. *Economic Geology*, **58** (8), 1246–1266.

Schuurmans, J. M., Bierkens, M. F. P., Pebesma, E. J. & Uijlenhoet, R. 2007 Automatic prediction of high resolution daily rainfall fields for multiple extents: the potential of operational radar. *Journal of Hydrometeorology*, **8** (6), 1204–1224.

Sharifi, E., Saghafian, B. & Steinacker, R. 2019 Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques. *Journal of Geophysical Research: Atmospheres*, **124** (2), 789–805.

Suhaila, J. & Jemain, A. A. 2012 Spatial analysis of daily rainfall intensity and concentration index in Peninsular Malaysia. *Theoretical and Applied Climatology*, **108** (1-2) 235–245.

Tabios, G. Q. III & Salas, J. D. 1985 A comparative analysis of techniques for spatial interpolation of precipitation. *JAWRA Journal of the American Water Resources Association*, **21** (3), 65–380.

Teegavarapu, R. S. V & Chandramouli, V. 2005 Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, **312** (1), 191 – 206.

Webster, R. & Oliver, M. A. 1993 How large a sample is needed to estimate the regional variogram adequately? *Soares A. (eds) Geostatistics Tróia'92. Quantitative Geology and Geostatistics*, **5**.

Yang, X., Xie, X., Liu, D. L., Ji, F. & Wang, L. 2015 Spatial interpolation of daily rainfall data for local climate impact assessment over greater Sydney region. *Advances in Meteorology*.