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

Research article

Statistical Modeling of Daily Rainfall Using Zero-tweaked Data

Aneeta Kalor[1], Rhysa McNeil[1, 2*], and Nurin Dureh[1]

[1] Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, 94000, Thailand
[2] Center of Excellence in Mathematics, Commission on Higher Education (CHE), Ministry of Education, Ratchathewi, Bangkok, 10400, Thailand
* Correspondence: nchirtki@gmail..com

Citation:
Kalor, A.; McNeil, R.; Dureh, N. Statistical modeling of daily rainfall using zero-tweaked data. *ASEAN J. Sci. Tech. Report*. **2026**, *29*(1), e259596. https://doi.org/10.55164/ajstr.v29i1.259596.

Article history:
Received: May 31, 2025
Revised: September 16, 2025
Accepted: October 9, 2025
Available online: December 14, 2025

Abstract: The significant amount of zero rainfall led to a highly skewed distribution of rainfall, creating challenges in rainfall modeling. This study aims to introduce a zero-tweaking method for handling a large proportion of zero rainfall data and apply it to daily rainfall data collected at four stations in southern Thailand from 2010 to 2022. The fourth root transformation was used to handle the right skew of rainfall. Zero-tweaking techniques were employed, with zeros substituted by normally distributed random numbers that permitted negative values. The patterns and trends in rainfall at each of the four stations were investigated using natural cubic splines. The trend projection analysis for four rainfall stations up to 2030 revealed an increase in rainfall at two stations in the Gulf of Thailand; however, this increase was not statistically significant. However, this study introduced the zero-tweaking method to handle the zero data, which enabled the use of conventional statistical methods and enhanced the model's validity.

Keywords: Natural cubic spline; rainfall; seasonal pattern; zero-tweaked data

1. Introduction

Rainfall is a crucial climate component that significantly influences various aspects of life and ecosystems. Rainfall plays a vital role in the water cycle. Global rainfall patterns have become increasingly unpredictable due to climate change, heightening the risk of flooding [1-2], leading to both heavy rainfall and prolonged droughts [3-5], and threatening social stability, water resources, and food security [6]. In Thailand, rainfall variability has been recognized as a significant factor influencing agriculture. Annual and seasonal rainfall parameters exhibited heterogeneity, with both increasing and decreasing trends observed at the national and regional levels [7]. Southern Thailand is significantly impacted by the southwest monsoon, which often results in intense rainfall and severe flooding [8]. Therefore, understanding rainfall patterns is critical for effective hydrological planning and management.

Rainfall data often contains a large number of zeros, resulting in a highly skewed distribution. This is due to the nature of rainfall, where most days have little or no rainfall, while a few days experience heavy downpours. This feature provides challenges and opportunities for statistical analysis and modeling. Several non-normal distribution models were used to analyze the right-skewed distribution of the rainfall, such as the Gamma distribution model [9-12], the Exponential distribution model [13-14], and the Weibull distribution model [15]. Some studies recommend transforming the data using a log-normal distribution by taking the logarithm of the rainfall data [11]. Another issue to consider when analyzing rainfall data is the large number of zero rainfall events. The statistical models have been designed to handle the large amount of zero data, such as the

zero-inflated Poisson regression model, which is designed to handle count data with an excess of zero observations [16], and the hurdle model [17].

Other methods included two process of Poisson–Gamma approach, where the Poisson distribution represented the daily occurrence of rainfall events, and the Gamma distribution captured the intensity of those events [18]; a stochastic model, a two-state Markov chain model, to represent the occurrence of rainfall and then used other probability models, gamma and exponential, to fit the intensity of rainfall [19]. Some statistical models for addressing this issue include separating the rainfall data into a binary model (zero or non-zero) and a conditional model for non-zero data, or using zero-inflated models that assume zero values in the data arise from structural zeros or random zeros. The issue of excessive zeros has been addressed in limited research through the use of continuous random numbers for replacement. Therefore, the objective of this study is to address high skew and a large amount of zero rainfall data by using the fourth root transformation scale and zero-tweaking for the rainfall data in southern Thailand. After addressing the distribution and zero rainfall, this study also investigates rainfall patterns using a cubic spline model.

2. Materials and Methods

2.1 Data and Area of Study

Data used for this study were obtained from the National Oceanic and Atmospheric Administration (NOAA). It is available for public assessment, which can be downloaded from https://www.ncdc.noaa.gov. NOAA collected rainfall data from ground stations. This study considers the daily rainfall in southern Thailand. The southern region has a tropical monsoon climate. A long, pointed peninsula characterizes the southern region's geography. There is a surface of water flanking both the western side, the Andaman Sea, and the eastern side, the Gulf of Thailand, causing rain all year round and being the region with the most rainfall in Thailand. The rainfall data used in this study are from four stations in southern Thailand, where two stations, Phuket Airport and Trang, are located in the Andaman Sea, and two stations, Nakhon Si Thammarat and Hat Yai, are located in the Gulf of Thailand. We chose these stations because the southwest monsoon directly affects the Andaman coast, Phuket, and Trang. Meanwhile, the northeast monsoon has a significant impact on the coast of the Gulf of Thailand, particularly in Nakhon Si Thammarat and Hat Yai.

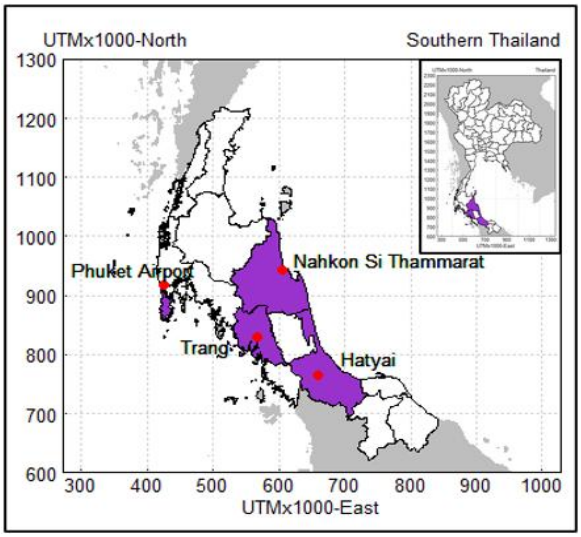

Figure 1. Locations of the study area

The data consisted of missing data and a large number of zeros. After managing the data, the raw dataset consisted of 4,748 daily records per station (totaling 18,992 records for four stations) during 2010–2022. Of these, 6,681 entries (35.18%) were missing, and 2,237 entries contained zero rainfall.

2.2 Statistical Methods

The fourth root transformation was used to handle the right skew of rainfall data and assess it using a quantile-quantile (Q-Q) plot. A comparison of the efficiency of data transformation methods using simulated

data found that the fourth root transformation demonstrated the best results. The authors then applied this transformation to the rainfall data [20]. Then, the zero-tweaked method was used to handle a large number of zero rainfall values by replacing all values with non-zero ones, allowing for negative numbers since the rainfall data was transformed using the fourth root scale; zero rainfall remained zero. This was done by splitting the transformed rainfall data into two separate sets using a value close to zero; we use the minimum number, which equals 0.74 mm, as the cut point. The first data set includes all data from the cut point that has a normal distribution after the fourth root transformation. The second data set includes all data below the cut point, most of which is zero. This second set of data was replaced with a normal random generation number, with a mean and standard deviation approximated from the first data set. This method enabled the transformed data to include negative numbers, allowing the data to achieve a normal distribution —a key assumption of conventional statistical methods. The normality was assessed using a Q-Q plot. Once the data are typically distributed, classic statistical methods can be used for further investigation.

The trends and seasonal patterns of rainfall at each station were displayed using a cubic spline function with a linear function. Cubic spline functions are defined as piecewise polynomials with degree n, where knots are the chosen positions that join the pieces. A spline function of degree n is a continuous function with $n-1$ continuous derivatives [21-22]. Using an adequate number of knots, a time series plot was created for each rainfall and station, and the positions of the knots were fixed to smooth the spline curve. The formula denoting the cubic spline model is given in Equation 1 as follows:

$$s_i = a + bt_i + \sum_{k=1}^{p} c_k \, (t_i - t_k)_+^3 \tag{1}$$

where s_i is the spline function, a, b, and c_k are model parameters, k is the knot location, t_i denotes time in eight days, that is specified from 13 years, $t_i < t_2 < \ldots < t_p$ are specified knots and $(t_i - t_k)_+$ implies that $(t_i - t_k)$ is positive for $(t_i - t_k)$ and zero otherwise. The number of knots used to smooth the data is important for cubic spline fitting. Some studies chose based on the season [23-24]. This study used eight knots based on season and tropical region characteristics. The rainfall data were then seasonally adjusted by subtracting the fitted values, obtained from the cubic splines, from observed rainfall, then adding back the overall mean of rainfall, using the formula in Equation 2 below:

$$z_i = y_i - s_i + \bar{y} \tag{2}$$

where, z_i is the seasonally adjusted rainfall at observation i, y_i is the rainfall observation, s_i is the fitted value from the spline model and $\bar{y}$ is the observed rainfall overall mean. Autoregression (AR) models were also used to account for the autocorrelations among the residuals from the fitted linear models [25]. The seasonally adjusted rainfall was then fitted using a first-order autoregressive model to account for the autocorrelation. Then, a simple linear regression model and a natural cubic spline model were used to investigate the pattern and trend of seasonally adjusted rainfall

3. Results and Discussion

The distribution of rainfall was first investigated by combining the data from four stations into a single dataset, as shown in Figure 2. Figures 2(a1) and 2(a2) displayed the histogram and normal Q–Q plot of the aggregated data from the four stations. The histogram showed a highly skewed distribution, while the Q–Q plot revealed significant departures from the reference line, particularly in the upper quantiles. Figure 2(b1) and 2(b2) employed a fourth root transformation, which displayed the distribution closer to normality but remained with a large number of zeros. It showed a gap between zero rainfall and the minimum rainfall value (0.74 mm). The transformation reduced skewness but did not fully achieve normality due to the presence of a large number of zero values. These zeros were handled by first replacing any numbers less than the minimum with a uniform distribution. This technique did not allow for negative values, as shown in Figures 3(a1) and 3(a2). The distribution of the fourth root transformation still departed from the normal distribution (Q-Q plot). Therefore, we applied the zero-tweaked method by replacing zero with non-zero values, allowing for negative values, as shown in Figures 3(b1) and 3(b2). This method effectively adjusts the data to align closely with a normal distribution.

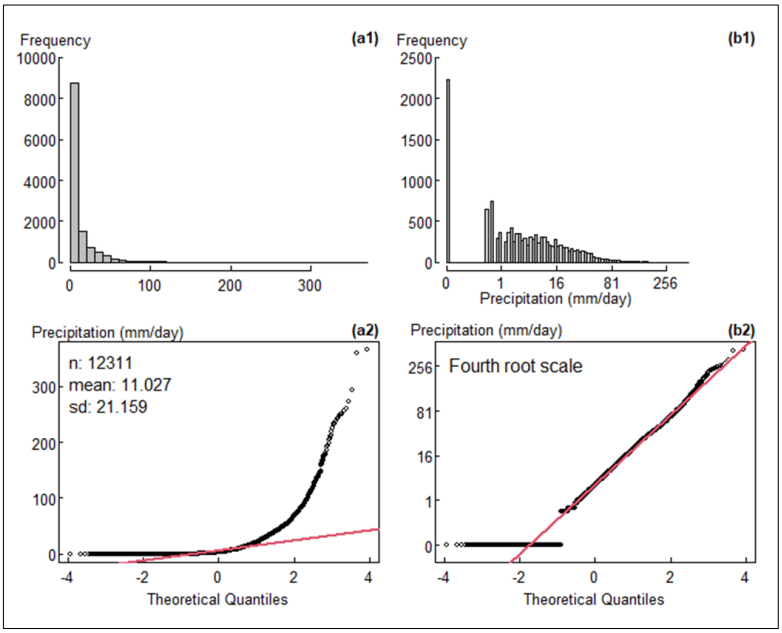

Figure 2. Distribution of fourth root of the daily rainfall

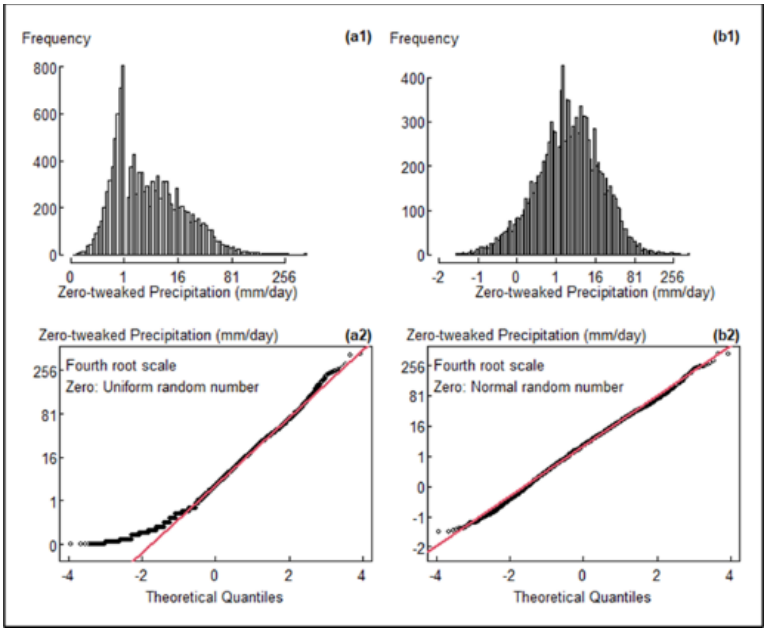

Figure 3. Distribution of the fourth root, replacing zero with a normal random number

Next, we investigate the pattern and trend for each station using the zero-tweaked fourth root ransformation of rainfall. Figure 4 shows the detailed steps of the analysis for the Nakhon Si Thammarat station.

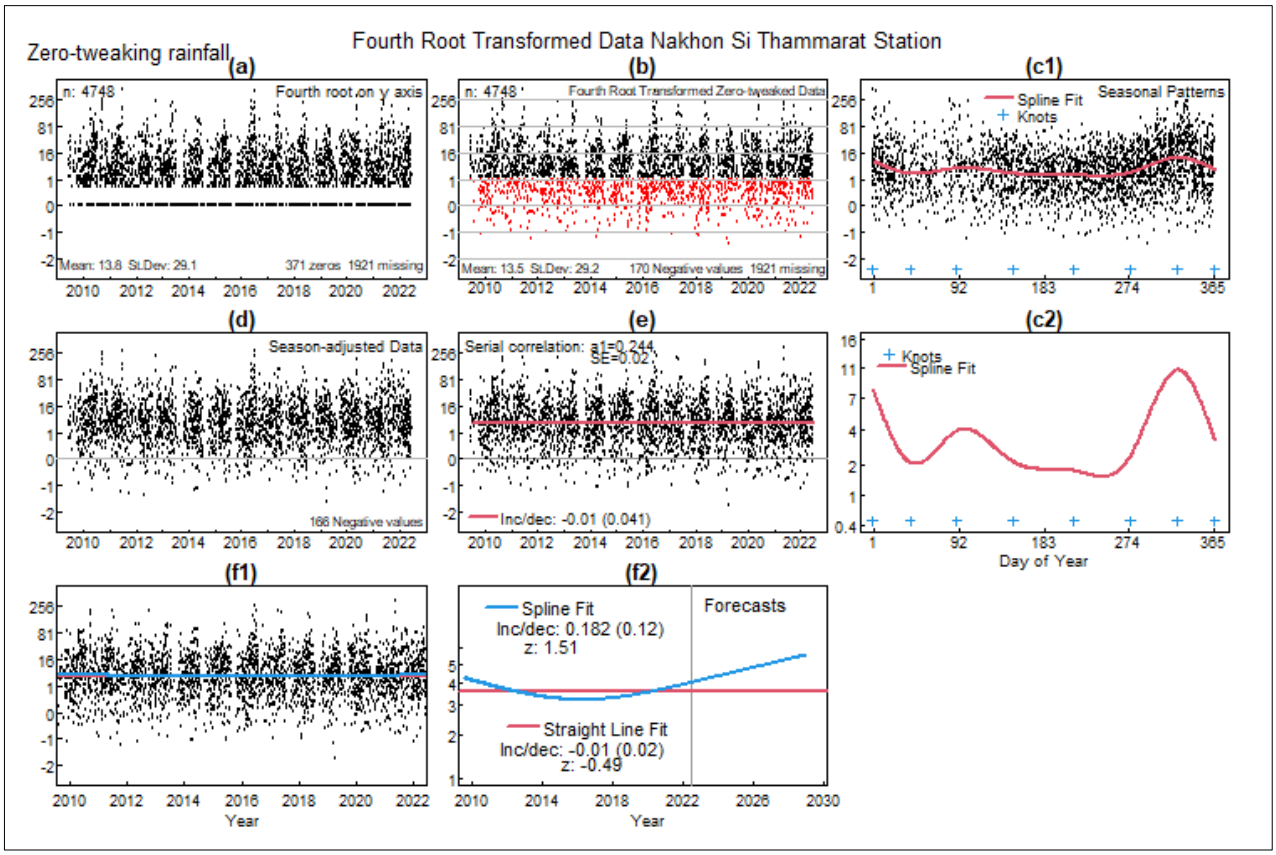

Figure 4. The pattern and trend of zero-tweaking rainfall for the Nakhon Si Thammarat station

Figure 4(a) displays the time series plot of rainfall data after transformation using the fourth root. Figure 4(b) displays the time series after applying the fourth root transformation and zero-tweaking to the data. All the red dots represented normal, random data replacing the zeros. Both graphs showed a seasonal pattern for each year. The natural cubic spline was used to fit the seasonal patterns shown by the red curve, with 8 knots denoted by blue crosses in Figure 4(c1). The fitted curve was then enlarged to illustrate the pattern clearly in Figure 4(c2). It clearly showed that the peak in rainfall occurred between the end of the year and the beginning of the year, with decreased rainfall in the middle of the year. The rainfall data was seasonally adjusted as shown in Figure 4(d). The first autoregressive model was used to remove the autocorrelation in Figure 4(e). The estimated coefficient of the autoregressive AR (1) model was small, indicating that the seasonally adjusted rainfall is independent.

The pattern and trend were fitted using a simple linear regression model (red curve) and a natural cubic spline (blue curve) in Figure 4(f1), and were enlarged to see the pattern clearly in Figure 4(f2). The results of the straight-line fit showed no trend; the spline fit, however, indicated a trend of decreasing values followed by a slight increase, with projections extending until 2030. However, all the absolute z-value statistics were less than 1.96, indicating that they were not statistically significantly different. The patterns and trends for the four stations were presented in Figure 5. Figure 5 (top panel) shows the seasonal patterns of the zero-tweaked fourth root transformation of rainfall after fitting a natural spline function with eight knots for all four stations, and Figure 5 (bottom panel) enlarges the scale. The results showed two seasonal patterns. Nakhon Si Thammarat and Hat Yai stations are located in the Gulf of Thailand. The peak in rainfall occurs at the end of the year, during October and November, which marks the rainy season. Phuket and Trang stations, situated on the Andaman Sea, experienced the highest rainfall from June to September. Therefore, the zero-tweaked fourth root transformation was seasonally adjusted. The first autoregressive model was also used to remove autocorrelation for all stations.

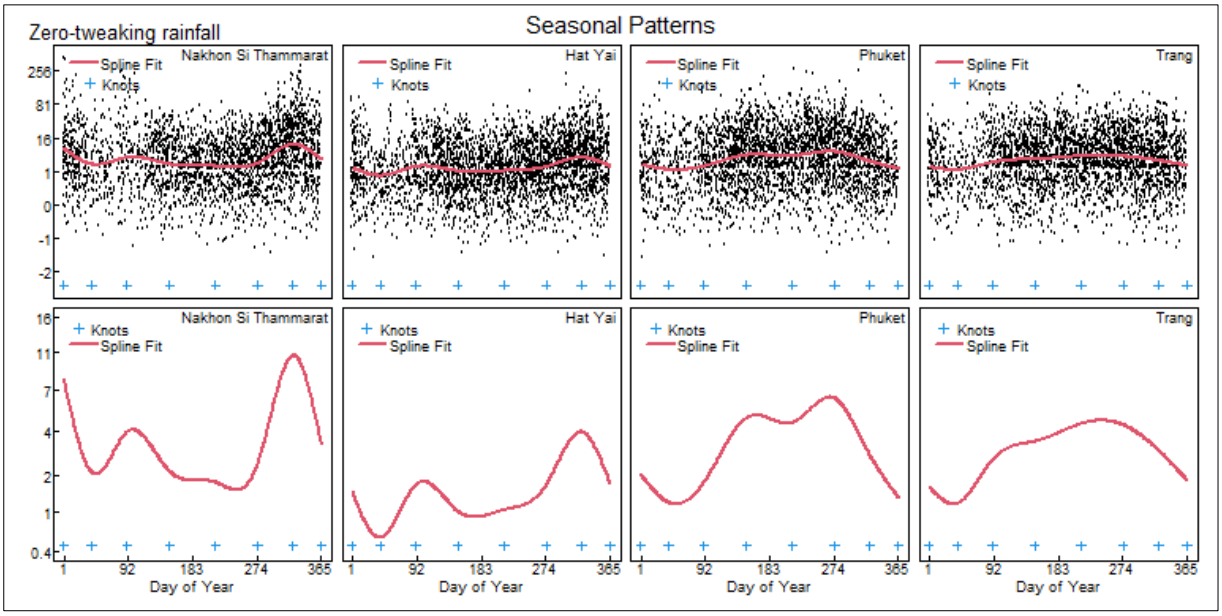

Figure 5. Seasonal patterns of zero-tweaking rainfall for four stations

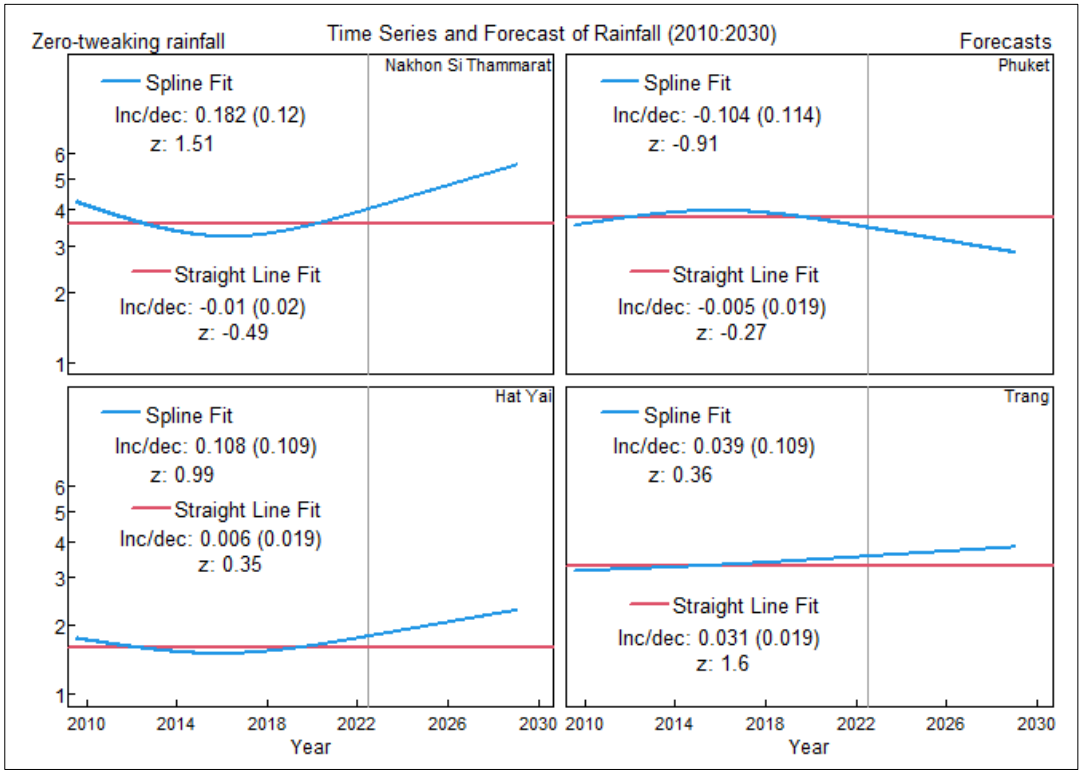

Figure 6. Patterns and forecasting trend of zero-tweaking rainfall in 2030 for four stations

Figure 6 illustrates the enlarged patterns, trends, and forecasting trends in 2030 of the zero-tweaked fourth root transformation using simple linear regression (straight red line) and the natural cubic spline (blue line) model for all four stations. This was examined after seasonally adjusting and removing autocorrelation. The results clearly showed no linear trend at any of the stations. However, the results from the natural cubic spline showed that the Nakhon Si Thammarat and Hat Yai stations had a decreasing trend from 2010 to 2016, followed by an increase up to 2022. For both stations, the forecasting trend increased in 2030. Whereas Phuket stations experienced an upward trend from 2010 to 2016, they then gradually declined afterward. Trang station showed a steady increasing trend. The pattern was consistent with a prior study on rainfall variability and

trends in Thailand, which found that the southern regions had an increase in rainfall, with an annual rainfall of 0.148 mm [26]. However, the z-value statistics displayed in the figure are based on statistical testing; there is no average rainfall change in each station. In this study, we transformed the right-skewed rainfall using a fourth root transformation, resulting in rainfall data that closely resembled a normal distribution. A previous study, which compared the efficiency of data transformation methods using simulated data, found that the fourth root transformation yielded the best results. The authors then applied this transformation to the rainfall data [20]. The zero rainfall remained zero after the fourth root transformation. In this study, we replaced the zeros with continuous random numbers from a normal distribution, allowing for small negative values. A previous study also employed the method of substituting all zero rainfall values with small negative values to identify dry areas in spatial interpolations of rainfall events, rather than leaving them as zero [27].

4. Conclusions

This zero-tweaked technique is straightforward to use and provides significant advantages in statistical modeling and data analysis. It created a smoother and more acceptable distribution, with the small negative values of the data still maintaining the relative difference between dry and wet conditions. The data can be more closely approximated to a normal distribution, which improves the validity of parametric tests and models. The zero-tweaked rainfall at four stations was then seasonally adjusted. The patterns and trends were investigated using a natural cubic spline regression model. Nakhon Si Thammarat and Hat Yai stations, located in the Gulf of Thailand, exhibited a decreasing trend until 2016, followed by an increasing trend, with a projected continuation of the latter trend until 2030. In contrast, Phuket station exhibited an increasing trend until 2016, followed by a decreasing trend, with a projected decrease continuing until 2030. However, the statistical tests for the changes in our study were not significant. Additionally, the accuracy of long-term forecasts should be evaluated. The rainfall also included a significant amount of missing data, which requires management. This study applied the methods to only four selected stations; these patterns may not represent statistically significant changes. Further investigation with additional rainfall stations in southern Thailand is recommended.

5. Acknowledgments

The authors sincerely appreciate Professor Don McNeil for his invaluable assistance. The Faculty of Science and Technology, Pattani Campus, Thailand, supported this work..

Author Contributions: "Conceptualization, A.K., R.M., and N.D.; methodology, R.M.; data curation and data analysis, A.K. and N.D.; writing draft preparation, A.K. and R.M.; review and editing, R.M. and N.D.. All authors have read and agreed to the published version of the manuscript.

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

References

[1] Shu, E.G.; Porter, J.R.; Hauer, M.E.; Olascoaga, S.S.; Gourevitch, J.; Wilson, B.; Pope, M.; Vazquez, D.M.; Kearns, E. Integrating climate change induced flood risk into future population projections. Nat. Commun. **2023**, *14*, 7870. https://doi.org/10.1038/s41467-023-43493-8

[2] Tabari, H. Climate change impact on flood and extreme precipitation increases with water availability, Sci. Rep. **2020**, *10*(1), 13768. https://doi.org/10.1038/s41598-020-70816-2

[3] Bevacqua, E.; Rakovec, O.; Schumacher, D.L.; Kumar, R.; Thober, S.; Samaniego, L.; Seneviratne, S.I.; Zscheischler, J. Direct and lagged climate change effects intensified the 2022 European drouth. Nat. Geosci. **2024**, *17*, 1100-1107. https://doi.org/10.1038/s41561-024-01559-2

[4] Cook, B.I.; Mankin, J.S.; Williams, A.P.; Marvel, K.D.; Smerdon, J.E.; Liu, H. Uncertainties, limits, and benefits of climate change mitigation for soil moisture drought in southwestern North America. *Earth's Future*, **2021**, *9*(9), e2021EF002014. https://doi.org/10.1029/2021EF002014

[5] Nielsen, M.; Cook, B.I.; Marvel, K.; Ting, M.; Smerdon, J.E. The changing influence of precipitation on soil moisture drought with warming in the Mediterranean and Western North America. *Earth's Future,* **2024**, *12,* e2023EF003987. https://doi.org/10.1029/2023EF003987

[6] Ide, T.; Fröhlich, C.; Donges, J.F. The economic, political, and social implications of environmental crises. *Bull. Am. Meteorol. Soc.* **2020**, *101*, E364–E367. https://doi.org/10.1175/BAMS-D-19-0257.1

[7] Madolli, M.J.; Gade, S.A.; Gupta, v.; Chakraborty, A.; Cha-um, S.; Datta, A.; Himashu, S.K. A systematic review on rainfall patterns of Thailand: Insights into variability and its relationship with ENSO and IOD. *Earth-Sci. Rev.* **2025**, *264,* 105102. https://doi.org/10.1016/j.earscirev.2025.105102

[8] Limsakul, A.; Limjirakan, S.; Sriburi, T. Observed changes in daily rainfall extremes along Thailand's coastal zone. *J. Environ. Res.* **2010**, *32*, 49-68.

[9] Owusu, B.E.; McNeil, N. Statistical modelling of daily rainfall variability patterns in Australia. Pertanika *J. Sci. Technol.* **2018**, *26*(2), 691-706.

[10] Vieira, F.M.C.; Machado, J.M.C.; Vismara, E.; Possenti, J.C. Probability distributions of frequency analysis of rainfall at the southwest region of Paraná State, Brazil. Rev. Cienc. Agroveterinarias **2018**, *17,* 260–266. https://doi.org/10.5965/223811711722018260

[11] Hasan, M.M.; Croke, B.F.W.; Liu, S.; Shimizu, K.; Karim, F. Using mixed probability distribution functions for modelling non-zero sub-daily rainfall in Australia. *Geosciences,* **2020**, *10*(2), 43. https://doi.org/10.3390/geosciences10020043

[12] Ximenes, P.; Silva, A.S.A.; Ashkar, F.; Stosic, T. Best-fit probability distribution models for monthly rainfall of Northeastern Brazil. Water Sci. Technol. **2021**, *84*(6), 1541-1556. https://doi.org/ 10.2166/wst.2021.304

[13] Sake, R.; Akhtar, P.M. Fitting of modified exponential model between rainfall and ground water levels: A case study. *Int. J. Stat. Appl. Math.* **2019**, *4*(4), 1-6.

[14] Afolabi, A.M.; Adesola, O.I. Exponential probability distribution of short-term rainfall intensity. Equity J. Sci. Technol. **2022**, *9*(2), 18-27.

[15] Olivera, S.; Heard, C. Increases in the extreme rainfall events: using the Weibull distribution. *Environmetrics* **2018**, *30*. https://doi.org/10.1002/env.2532

[16] Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics* **1992**, *3*(1), 1-14. https://doi.org/10.2307/1269547

[17] Feng, C.X. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications,* **2021**, *8*, https://doi.org/10.1186/s40488–021–00121–4

[18] Dzupire, N.C.; Ngare, P.; Odongo, L.A. A Poisson-gamma model for zero inflated rainfall data. *J. Probab. Stat.* **2018**, *1012647*, https://doi.org/10.1155/2018/1012647

[19] Wilks, D.S. Multisite generalization of a daily stochastic precipitation generation model. J. Hydrol. **1998,** *210*(4), 178-191. https://doi.org/10.1016/S0022-1694(98)00186-3

[20] Kaewprasert, T.; Khamkong, M.; Bookkamana, P.A. A comparison of data transformation methods of generalized exponential distribution and estimation of summer rainfall in Chiang Dao, Chiang Mai. *Burapha Sci. J.* **2017**, *22*(3), 385-396.

[21] Wahba, G. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics, **1990**; pp 1-161. https://doi.org/doi:10.1137/1.9781611970128

[22] Wold, S. Spline functions in data analysis, *Technometrics* **1974**, *16*(1), 1-11. http://www.jstor.org/wqazstable/1267485

[23] Wongsai, N.; Wongsai, S.; Huete, A.R. Annual seasonality extraction using the cubic spline function and decadal trend in temporal daytime MODIS LST data. *Remote Sens*. **2017**, *9*(12), https://doi.org /10.3390/rs9121254

[24] Lukas, M.A.; de Hoog, F.R.; Anderssen, R.S. Efficient algorithms for robust generalized cross-validation spline smoothing. *J. Comput. Appl. Math.* **2010**, *235*(1), 102–107. https://doi.org /10.1016/j.cam.2010.05.016

[25] Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*. Springer, Queensland, **2002**. https://doi.org/10.1007/978-0-387-21706-2

[26] Waqas, M.; Humphries, U.W.; Hlaing, P.T. Time series trend analysis and forecasting of climate variability using deep learning in Thailand. *Results Eng.* **2024**, *24*, 102997. https://doi.org/10.1016/j.rineng.2024.102997

[27] Lee, T.; Shin, J.Y. Latent negative precipitation for the delineation of a zero-precipitation area in spatial interpolations. *Sci. Rep.* **2021**, *11*(1), 20426. https://doi.org/10.1038/s41598-021-99888-4