applied sciences

MDPI

Article

Sequential Data Processing for IMERG Satellite Rainfall Comparison and Improvement Using LSTM and ADAM Optimizer

Seng Choon Toh [1,*], Sai Hin Lai [1,*], Majid Mirzaei [2], Eugene Zhen Xiang Soo [3] and Fang Yenn Teo [4]

[1] Department of Civil Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia
[2] Department of Environmental Science and Technology, University of Maryland, College Park, MD 20742, USA
[3] Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang 43000, Malaysia
[4] Faculty of Science and Engineering, University of Nottingham Malaysia, Semenyih 43500, Malaysia
* Correspondence: 17131973@siswa.um.edu.my (S.C.T.); laish@um.edu.my (S.H.L.)

Abstract: This study introduces a systematic methodology whereby different technologies were utilized to download, pre-process, and interactively compare the rainfall datasets from the Integrated Multi-Satellite Retrievals for Global Precipitation Mission (IMERG) satellite and rain gauges. To efficiently handle the large volume of data, we developed automated shell scripts for downloading IMERG data and storing it, along with rain gauge data, in a relational database system. Hypertext pre-processor (pHp) programs were built to visualize the result for better analysis. In this study, the performance of IMERG estimations over the east coast of Peninsular Malaysia for the duration of 10 years (2011–2020) against rain gauge observation data is evaluated. Moreover, this study aimed to improve the daily IMERG estimations with long short-term memory (LSTM) developed with Python. Findings show that the LSTM with Adaptive Moment Estimation (ADAM) optimizer trained against the mean square error (*MSE*) loss enhances the accuracy of satellite estimations. At the point-to-pixel scale, the correlation between satellite estimations and ground observations was increased by about 15%. The bias was reduced by 81–118%, *MAE* was reduced by 18–59%, the root-mean-square error (*RMSE*) was reduced by 1–66%, and the Kling–Gupta efficiency (*KGE*) was increased by approximately 200%. The approach developed in this study establishes a comprehensive and scalable data processing and analysis pipeline that can be applied to diverse datasets and regions encountering similar domain-specific challenges.

Keywords: IMERG satellite rainfall; SQL relational database; LSTM; ADAM; Python; pHp

Citation: Toh, S.C.; Lai, S.H.; Mirzaei, M.; Soo, E.Z.X.; Teo, F.Y. Sequential Data Processing for IMERG Satellite Rainfall Comparison and Improvement Using LSTM and ADAM Optimizer. *Appl. Sci.* **2023**, *13*, 7237. https://doi.org/10.3390/app13127237

Academic Editor: Cheng-Yu Ku

Received: 27 April 2023
Revised: 14 June 2023
Accepted: 15 June 2023
Published: 17 June 2023

1. Introduction

Precipitation recycling contributes to hydrological processes in the atmospheric branch of the water cycle [1]. Though the precipitation at the Earth's surface is typically measured with rain gauges [2], the coverage is surprisingly small [3]. Remote sensing applications have recently emerged as one of the most important methods of acquiring information on the Earth's surfaces [4]. For example, satellite precipitation estimation (SPE) products have played an important role in estimating the rainfall from the atmosphere [5–7]. Compared to traditional ground-based rain gauges and weather radars, the SPEs have significant advantages in terms of spatial coverage. Furthermore, SPEs are continuous and uniform, avoiding the high costs of ground observation networks [8] and could potentially be used for flood monitoring in rain gauge scarce areas [9,10].

Nowadays, many SPEs have been developed to satisfy various hydrometeorological needs [11], including the Tropical Rainfall Measuring Mission (TRMM) of the Multi-satellite Precipitation Analysis (TMPA) products [12], Climate Prediction Center Morphing

(CMORPH) [13], the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) [14,15], the Global Satellite Mapping of Precipitation (GSMaP) [16], Climate Hazards Group Infrared Precipitation with Station (CHIRPS) data [17], and the Integrated Multi-Satellite Retrievals Global (IMERG) precipitation mission [18]. Many studies on assessing the performance and accuracy of these SPEs have been made over the world [19–25]. Although these spatial-temporal datasets are now freely available, employing them is a great challenge as they are too large and complex. Although research has been performed to design data series management [26], researchers and hydrologists are still facing great challenges such as computational requirements as well as software and data storage issues when handling this huge amount of data, which is hindering its accessibility [27]. It is challenging to deal with a bigger scale of data in satellite and gauge precipitation research [28]. In addition, the acquisition, searching, transfer, analysis, and visualization of the data in many areas, such as geoscience, remote sensing, hydrology, and environmental research, are also challenging. Google Earth Engine's cloud computing platform has effectively addressed the challenges of big data analysis. However, there are still several main limitations, such as privacy, tool restrictions, selected data mining models limitations, features limitations, computational restrictions, third-party software causing time-consuming processes, and others [29].

The accuracy and reliability of SPEs are also among the concerns of hydrologists [30]. The high spatial-temporal datasets can be biased depending on the target terrain, elevation, and season. Therefore, many researchers have dedicated their efforts to exploring the application of machine learning (ML) and deep learning (DL) in improving the accuracy and reliability of SPEs [31–33]. Recurrent neural network (RNN) is a deep learning approach for modeling sequential data and has been widely used in different types of applications such as speech recognition, natural language processing, energy management, and time series analysis [34–37]. Long short-term memory (LSTM), an RNN architecture used in the field of DL [38,39], is one of the most popular methods to be applied to improving SPEs [40–43]. It was proposed by Hochreiter and Schmidhuber in 1997 [43]. In a recent study conducted by Moazam et al. (2023), various deep learning methodologies were examined to assess their effectiveness in predicting the streamflow for the Muda River Basin which is located in the northern part of Peninsular Malaysia. The findings indicate that LSTM outperformed other approaches, demonstrating superior performance [44]. Our study attempts to improve the SPEs by merging the satellite rainfall and gauge precipitation using the LSTM application with the adaptive moment estimation (ADAM) optimizer. ADAM optimizer has shown good results in deep learning modeling [45]. Research performed in China to test the prediction capabilities of the weighted mean temperature, LSTM outperformed traditional RNN [46]. Generally, LSTM utilizes certain types of artificial memory processes that can help more effectively imitate human thought. LSTM works in the manner that, with discrete time steps t = 0, 1, 2, 3, 4, 5, 6, …, units' activation will be updated (forward pass) followed by all weights' error signals computation (backward pass) [47]. Previous research has shown that LSTM is able to address the vanishing gradient problem when learning long-term dependencies that arise in traditional RNNs [48]. It has been proven that this method is better than the traditional RNN and is capable of learning long-term dependencies much faster [40,43,49,50]. The LSTM used in our study is one of the variants in LSTM architecture that adds "peephole connections" to let the gate layers look at the cell state to learn the precise timing of the outputs [47].

Managing large volumes of data poses a significant challenge for researchers in various fields. To address this issue, we propose a sequential program that aims to fill the gap and provide an effective solution for handling massive amounts of data. In this paper, we present the design and implementation of the program, along with its performance evaluation, comparison, and enhancement of the SPE. A systematic methodology is introduced to perform the data extraction and comparison of the satellite rainfall estimations with rain gauge observations with a sequence of programs using different technologies, such as Python for statistical analysis and modeling, pHp for the interactive visualizations of the

results, MariaDB for relational database storage with SQL for data pre-processing, and Shell Script for data extraction automation. According to a recent study conducted by Roh et al. (2021) concerning the data collection survey for machine learning, the importance of data cleaning due to the prevalent presence of data noise was highlighted [51]. It is essential to identify and remove erroneous data points from both satellite rainfall estimations and rain gauge observations [52,53] to enhance the quality of the data. SQL programs are used to filter the erroneous dataset. Data visualization in the form of a scatter plot and time series graph could help in effectively comparing the performance of the dataset [54]. In this paper, we propose a novel approach to develop a comprehensive and scalable data processing and analysis pipeline for the deep learning modeling of rainfall data that can be applicable and adaptable to different datasets and regions. The proposed approach can be effectively applied to various datasets and regions, as long as comparable resources such as satellite rainfall and rain gauge data are accessible and the procedures outlined in this investigation are followed. Additionally, the same approach can be applied to forecasting and modeling in other fields of study.

Researchers have been using rain gauge as the reference to enhance the satellite rainfall data using machine learning. A research was performed to improve the IMERG product with ML over the Brahmaputra River Basin, which used rain gauge observations as a reference [55]. Similarly, another study was performed to improve the satellite rainfall estimation from MSG data in northern Algeria with machine learning using rain gauge observation as reference [56]. The framework of this study was applied to the east coast of Peninsular Malaysia for generating daily precipitation estimates based on the rain gauge observations and IMERG satellite estimations, and the performance before and after enhancement is evaluated. The details of this methodology are explained in the next section. The findings are then discussed, and the final section concludes this research.

2. Materials and Methods

2.1. Study Site

The study site on the east coast of Peninsular Malaysia, comprised three states, namely the Pahang, Terengganu, and Kelantan states, which were selected for the case study. This study site is located on the eastern side of Peninsular Malaysia and is bordered by the South China Sea to the east. It has a unique topography that influences the rainfall pattern in the region. The land usage patterns for this study site vary across different categories, including residential areas, agricultural land, natural reserves, industrial zones, and commercial development. This region was mainly selected based on its great history of flood and high rainfall variability due to the monsoon cycle, and is thus in need of flood mitigation and water resource management. This study site experiences two monsoon seasons, the northeast monsoon from November to March and the southwest monsoon from May to September. During the northeast monsoon, the region might be experiencing heavy rainfall, while the southwest monsoon brings drier conditions to the region. This study showed that there were rainfall-related extreme events during the monsoon periods for the east coast of Peninsular Malaysia [57–59]. Among all the rain gauge data downloaded from DID, there are 184 rain gauge stations for the study site, encompassing the region within 2°30′0″–6°30′0″ N and 101°–104° E being selected for this study. These rain gauge stations are selected due to the data completeness and the uniformity of the location within the study area. The rain gauge stations are shown in terrain view of Google Maps [60] as in Figure 1. In their research, Hu et al. [61] sought to define and measure terrain information from digital elevation model (DEM) to provide clear information for the various patterns of the land surface. It is noticed that the rain gauge station located at the boundary of Negeri Sembilan is included here due to the record as a rain gauge station in the state of Pahang. The mountainous regions located in the central part of the study area, particularly in the state of Pahang, have a lower density of rain gauge stations compared to the more residential areas in the region. For the deep learning training process, a total of 131 IMERG grids encompassing 184 rain gauge stations were utilized during the period 2011–2020.

The 184 rain gauge stations are selected based on their location in the state of Pahang, Terengganu, and Kelantan. With the longitude and latitude of the rain gauge stations, the 184 rain gauge stations are mapped into the 131 IMERG grids. If there is more than 1 rain gauge station in the same grid, the rain gauge data will be averaged. The IMERG data are then trained with the rain gauge data using the LSTM model developed in Python.

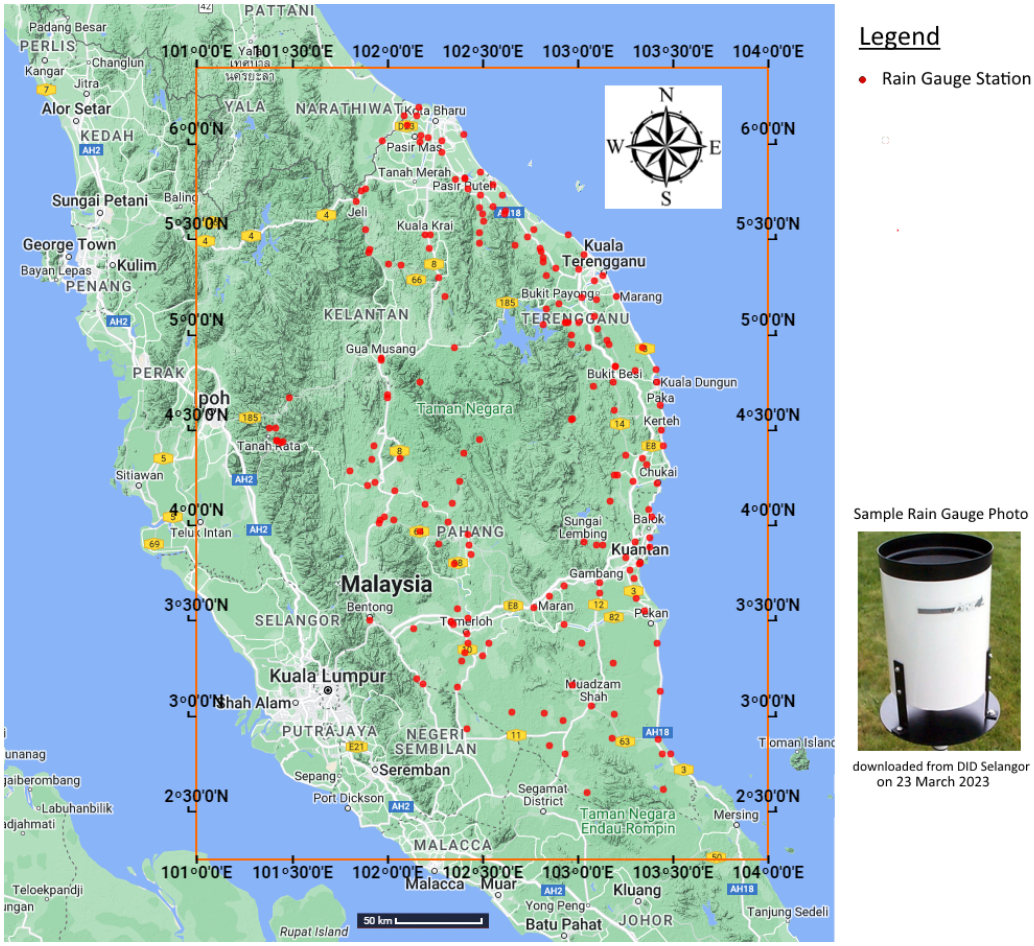

Figure 1. Peninsular Malaysia (PM) and the distribution of rain gauge observations on the east coast of PM and a sample rain gauge photo.

2.2. Methodology

The present study used an open source relational database system running on Linux to overcome the difficulty of handling a huge amount of data. Databases and tables are created separately to store the spatial-temporal datasets and they are retrieved from the query joins. The databases which were constructed using various technologies can be used in any programming tool for further studies including but not limited to implementing machine learning or deep learning models, as shown in Figure 2. It serves as a container to perform modeling. These databases can be linked or exported with the use of database connectors to facilitate their utilization in neural networks, machine learning, and deep learning modeling. Alongside conventional tools such as Matlab, the Python programming language has gained significant popularity in this domain due to its rich ecosystems and robustness, encompassing valuable libraries and frameworks such as Pytorch, Keras, and TensorFlow. The presence of these tools has greatly facilitated the process of AI modeling. Generally, this study has been divided into six phases: data acquisition, data retrieval, data import and storage, pre-processing, validation, and improvement. The details of each phase are explained in the following subsections. Figure 3 presents the overall methodology used in this study.

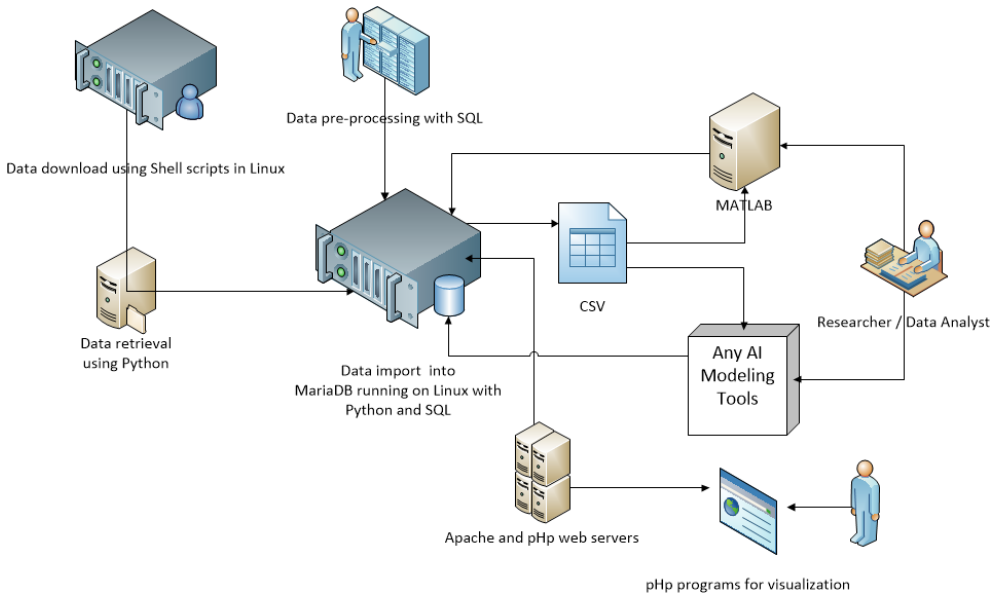

Figure 2. Schematic diagram of the relational database constructed using various technologies in connection with further research.

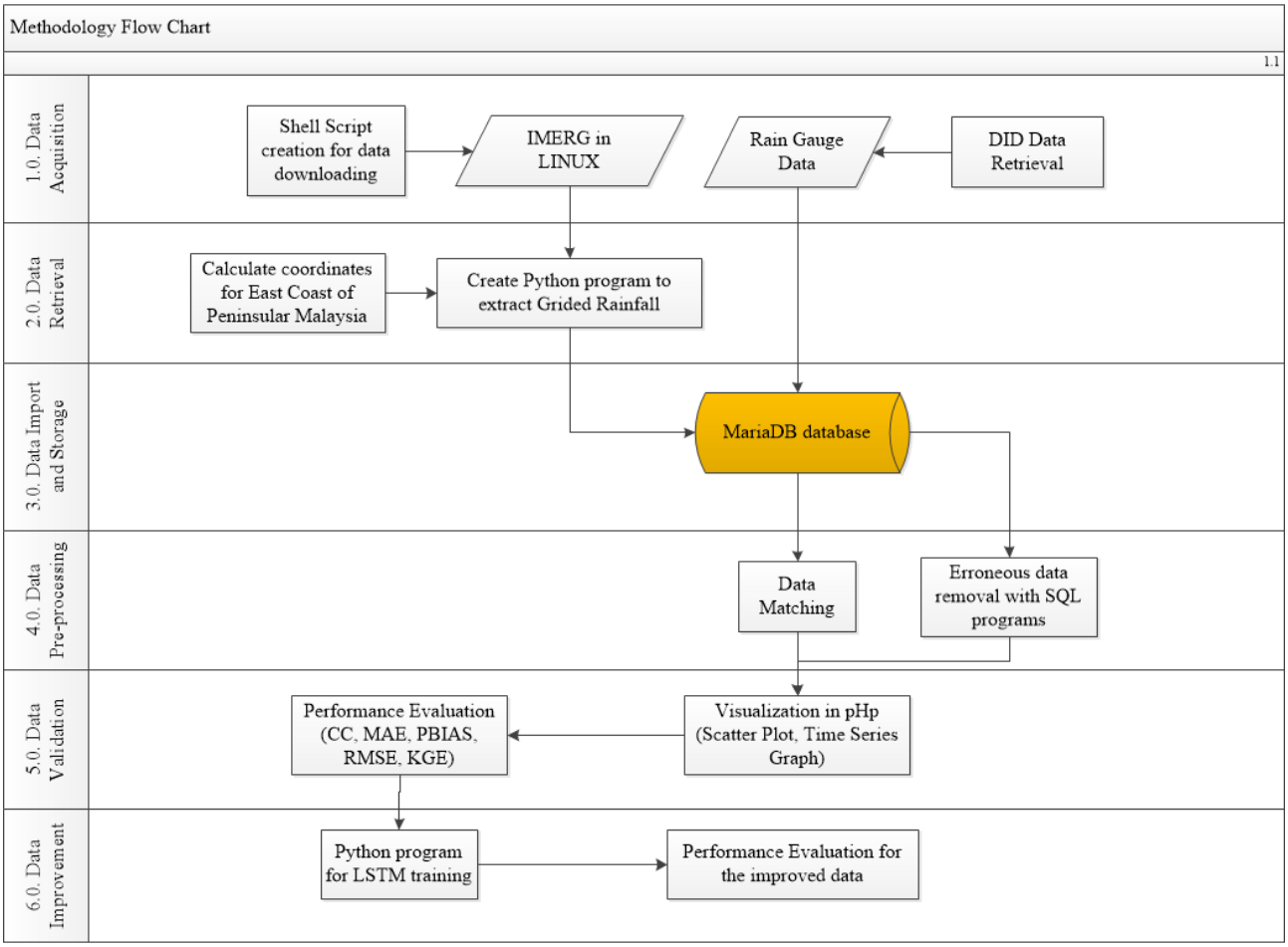

Figure 3. Methodology flow chart.

2.2.1. Phase 1: Data Acquisition

Rain gauge observations were acquired from the Department of Irrigation and Drainage (DID), Malaysia. The data were used as a comparison and for training purposes for the

satellite rainfall data. The distribution of rain gauge observations is presented in Figure 1. Satellite rainfall data can be acquired from the respective agencies' webpages. For example, TRMM and IMERG satellite files can be acquired from the Goddard Earth Sciences Data and Information Services Centre (GES DISC) (https://disc.gsfc.nasa.gov/ (accessed on 15 June 2021)), and PERSIANN satellite files can be obtained from the portal of the Center for Hydrometeorology and Remote Sensing (CHRS) (https://chrsdata.eng.uci.edu/ (accessed on 30 September 2021)). Before the satellite files were downloaded, CentOS Linux (https://www.centos.org/ (accessed on 10 June 2021)) was set up. CentOS is a freely available computing platform that serves as an ideal environment for executing command-line utilities such as the Wget computer program, which enables the seamless retrieval of files from the Internet. As it is open source, and community-supported, CentOS Linux is a rich base platform that provides a consistent, manageable platform for various deployments. Due to the vast number of files involved, we developed Shell Scripts incorporating the Wget program to facilitate the retrieval of IMERG satellite files in a batch processing manner. These IMERG satellite files are in Hierarchical Data Format version 5 (HDF5) format which is the standard mechanism for storing large quantities of numerical data [62].

The current research focuses on the Integrated Multi-Satellite Retrievals for Global (IMERG) precipitation measurement [18]. IMERG is an estimation that combines data from the Global Precipitation Measurement (GPM) satellite constellation to predict precipitation, which is particularly important over the majority of the Earth's surface that lacks precipitation-measuring instruments on the ground. Lately, with the IMERG Version 6, the algorithm merged the TRMM satellite's operation (2000–2015) with the GPM satellite (2014–present). These products have improvements in terms of coverage (60° N–60° S) as well as spatial (0.1°) and temporal (30 min) resolutions. In addition to the higher spatial-temporal resolution, IMERG provides three runs which are IMERG Early (IMERG-E), IMERG Late (IMERG-L), and IMERG Final (IMERG-F), to accommodate various consumer requirements for latency and accuracy. For the present study, the IMERG-F 30 min interval product at a spatial resolution of 0.1° was employed as it had been bias-corrected using the Global Precipitation Climatology Centre (GPCC) precipitation gauges and is suitable for scientific research.

2.2.2. Phase 2: Data Retrieval

Python Version 3.0 [63] is used to extract the data from IMERG. The satellite files downloaded contain a huge matrix of data for the globe. To extract the data of this study area we want, the boundary coordinate of the country or this study area has to be calculated based on the coverage that the satellite captured and its spatial resolution. As the spatial resolution of IMERG is 0.1 degree and has a spatial coverage of $[-180, 180]$ for longitude and $[-90, 90]$ for latitude, the IMERG grid calculation for an area can be derived as follows:

$$Longitude : (X)Grid = (X + 179.95)/0.1 \quad (1)$$

$$Latitude : (Y)Grid = (Y + 89.95)/0.1 \quad (2)$$

Initially, the whole Peninsular Malaysia boundaries are marked as rectangle size and defined with the longitude of 99.35 ($X1$) to 104.95 ($X2$) and the latitude of 1.05 ($Y1$) to 7.05 ($Y2$). With Equations (1) and (2), the IMERG grids are derived in matrix format as "$X1$ grid": "$X2$ grid"; "$Y1$ grid": "$Y2$ grid". Later, a Python program that makes use of h5py library was prepared to extract all IMERG grids that match the matrix for the whole of Peninsular Malaysia from the HDF5 data files. Due to the huge amount of data to be extracted, an EXE program was compiled from Python and a batch job was created to run continuously on the server. Hourly rain gauge observation records were obtained (in CSV format) from ground gauging stations from the DID.

2.2.3. Phase 3: Data Import and Storage

The extracted satellite data and the rain gauge data are imported into MariaDB. MariaDB is one of the open source relational databases. Although this database originated from the developers of MySQL, it has better performance in average import time [64]. The MariaDB Server code base is ensured to remain open for usage and contributions on technical merits (https://mariadb.org/ (accessed on 12 June 2021)). The MariaDB is being hosted in CentOS Linux. As compared with Microsoft Excel, MariaDB can handle a huge amount of data and there is no limitation problem faced when pumping the data into it. Indexing is performed for the tables for the fastest retrieval and the joining of the tables. Research showed that MariaDB has improved compression performance for flash devices, storage efficiency, and CPU utilization [65]. All the extracted IMERG data are transformed from HDF5 files into datasets in the tables in the databases with Python program. CSV files retrieved from DID are imported into tables using the data importing tool. The data can be retrieved anytime by anyone with a proper connection setup.

2.2.4. Phase 4: Data Pre-Processing

Structured query language (SQL) is written to filter the missing data in the database. SQL is a powerful tool for accessing and manipulating a relational database system [66]. The data pre-processing can be performed with Stored Procedures, which can be recalled for new datasets or re-processing of the data. SQL scripting files can be saved and called in the Linux terminal for a huge amount of data processing. The rain gauge data retrieved from DID were massaged to match the satellite data according to the latitude and longitudeof the study site.

If there is more than one rain gauge station in the same grid, the rain gauge values will be averaged. A simple SQL selection statement will tell us which grids have more than one rain gauge station. SQL is also used to determine which stations have complete data for the study period. Rain gauge stations that do not have complete data or missing data are excluded from the study. A preliminary examination of the rain gauge observation data using SQL selection indicates the presence of erroneous entries containing special characters such as # and $*$ in the rainfall column. These erroneous data were ignored by removing them from the datasets using SQL execution. The same spatial-temporal IMERG and rain gauge data were mapped using latitude and longitude values and stored in a table for the ease of retrieval using SQL. These data were normalized to obtain daily IMERG rainfall estimates by grid with rain gauge data for the study area for the period from 2011 to 2020.

2.2.5. Phase 5: Data Validation

The performance of the extracted satellite data was validated by comparing it with the in situ ground observations (rain gauge) at the point-to-pixel scale. Five statistical measures, including the coefficient of correlation (CC), mean absolute error (MAE), percent bias ($PBias$), root-mean-square error ($RMSE$), and the Kling–Gupta efficiency (KGE), were used in the validation. These metrics are commonly used in assessing the accuracy and reliability of rainfall estimation models [67,68]. Equations (3)–(7) show the aforementioned statistical measures, where S and G represent satellite/gridded and gauge precipitation, respectively, and n is the total number of measurements, i is the index of data, $\underline{S}$ is the average value of S_i, and $\underline{G}$ is the average value of G_i.

CC (also known as Pearson's correlation coefficient) is a measure used to assess the strength of the association between two variables, indicating the degree of interconnectedness between them [69]. It ranges between -1 and 1, where -1 represents a perfect negative correlation and 1 represents a perfect positive correlation. The sign of the CC indicates the direction of the trend, while the absolute value indicates the extent to which the relationship can be modeled linearly. A CC value of 0 implies no relationship between the variables. A CC value of 1 signifies a perfect positive relationship, suggesting that a

change in one variable leads to a corresponding change in the other variable in the same direction and with the same magnitude.

The *PBias* (also known as relative bias), is a statistical measure that quantifies the direction and magnitude of the bias in an estimator or a model. This measure has been applied in many data comparison studies [70,71]. It is calculated by comparing the average difference between the estimated values and the true values to the true values themselves, expressed as a percentage. If $PBias = 0$, the estimator or model is unbiased and has no systematic deviation from the true values. A positive *PBias* indicates that the estimator or model tends to overestimate the true values, while a negative *PBias* indicates underestimation.

MAE represents the average absolute deviation of the predictions from the true value [72,73]. Smaller *MAE* values indicate a better predictive performance, with zero representing a perfect prediction. On the other hand, *RMSE* is the standard deviation of the prediction errors [74]. The *RMSE* compares the difference between the estimated and observed values. The best approach computed the lowest value of the *RMSE*. Both the *MAE* and *RMSE* can be used to diagnose the variation in the errors in a set of forecasts. Usually, the computed *RMSE* is larger or equal to the *MAE*; the greater difference between them, the greater the variance in the individual errors in the sample. If the $RMSE = MAE$, then all the errors are of the same magnitude.

Kling–Gupta efficiency (*KGE*) is a multi-component performance metric. This goodness-of-fit measure was developed by Gupta et al. [75] to provide a diagnostically interesting decomposition of the Nash–Sutcliffe efficiency, which facilitates the analysis of the relative importance of its different components (correlation, bias, and variability) in the context of hydrological modeling. Later, Kling et al. [76] proposed a revised version of this index, to ensure that the bias and variability ratios are not cross-correlated. Equation (7) shows the formula of *KGE*, where r is the linear correlation coefficient, α is a measure of relative variability in the estimated and observed rainfall values, and β is the ratio between the mean estimated and mean observed rainfalls.

$$cc = \frac{\sum_{i=1}^{n}(G_i - \underline{G}) \times (S_i - \underline{S})}{\sqrt{\sum_{i=1}^{n}(G_i - \underline{G})^2} \times \sqrt{\sum_{i=1}^{n}(S_i - \underline{S})^2}} \tag{3}$$

$$PBias = \left(\frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n} G_i} - 1\right) \times 100\% \tag{4}$$

$$MAE = \frac{\sum_{i=1}^{n} |S_i - G_i|}{n} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(S_i - G_i)^2}{n}} \tag{6}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{7}$$

2.2.6. Phase 6: Data Improvement

As the SPEs are prone to errors, it is crucial to enhance the estimations before they are ready for any hydrological applications. Therefore, in the present study, the IMERG satellite data are trained based on rain gauge data using the LSTM model. The standard LSTM architecture has memory blocks in the recurrent hidden layer that contain gates to control the information flow. The LSTM algorithm used in this research is the variant with "peephole connections" that let the gate layers consider the cell state, as shown in Figure 4. In the LSTM algorithm, there are three types of gates, which are the forgotten gate, input gate, and output gate. These gates are the memory cells that remember the state in the network. The flow of input and output activations will be controlled by the input gate and the output gate, respectively [77]. Firstly, the forget gate will reset the cell states from the beginning and scale the internal state of the cell before adding to the cell input. The

sigmoid layer is used in the forget gate to make the decision. The next step contains the sigmoid layer which is the input gate layer that decides what values to update and the tanh layer that creates a vector of new candidate values. The new cell state was updated from the old cell state by considering what to forget and what new information to store with the time. The output will be based on the cell state by filtering with the sigmoid layer that decides which part of the cell state to output, put the cell state through tanh, and multiply it by the output of the sigmoid gate. Equations (8)–(13) were used to compute a mapping from the input sequence $x = (x_i, \ldots, x_T)$ to the output sequence $y = (y_i, \ldots, y_T)$ iteratively with time) t = 1 to T for the LSTM algorithm used.

$$i_t = \sigma(W_{ic}c_{t-1} + W_{ih}h_{t-1} + W_{ix}x_t + b_i) \tag{8}$$

$$f_t = \sigma(W_{fc}c_{t-1} + W_{fh}h_{t-1} + W_{fx}x_t + b_f) \tag{9}$$

$$o_t = \sigma(W_{oc}c_t + W_{oh}h_{t-1} + W_{ox}x_t + b_o) \tag{10}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{ch}h_{t-1} + W_{cx}x_t + b_c) \tag{11}$$

$$h_t = o_t \odot \tanh(c_t) \tag{12}$$

$$y_t = W_{yh}h_t + b_y \tag{13}$$

W is the matrix weight for the gates (i for input gate, f for forget gate, o for output gate). c and h are, respectively, the cell activation vector and cell output activation vector. σ is the sigmoid function, whereas $\odot$ is the Hadamad product.

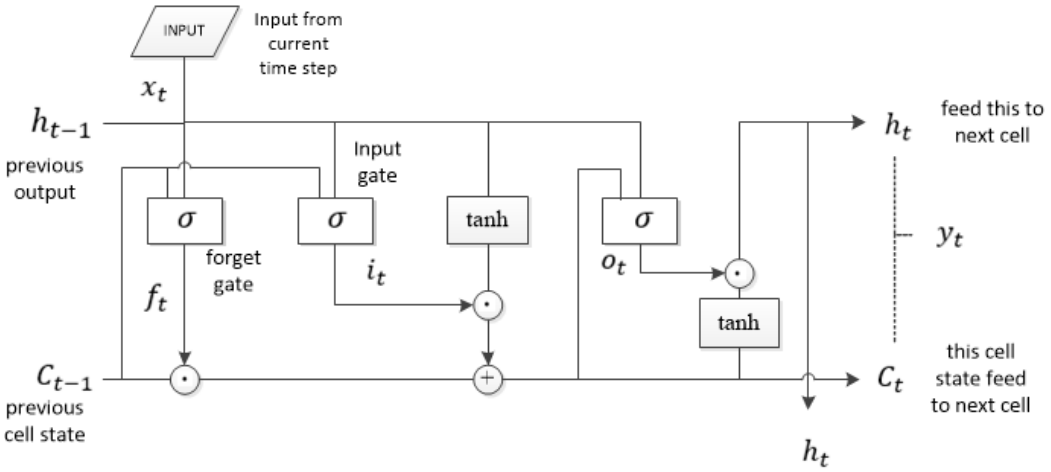

Figure 4. Proposed LSTM algorithm.

The proposed LSTM model was developed in Python that utilizes Pytorch to train IMERG data based on observations of rain gauge data. For grids that contain complete satellite and rain gauge data (10 years), the data are merged as per IMERG grids and divided whereby 2/3 of the data were used for training and the remaining 1/3 were used for validation purposes. IMERG and rain gauge data are loaded into tensors which constitute the Numpy alike n-dimensional array. The tensors are commonly used in deep learning modeling. In the single layer model, a sliding window approach is employed, where a sequence length of four IMERG records is utilized to predict a single rainfall record. The prediction is made with reference to the available rain gauge data. The daily IMERG data were trained for each grid with an epoch of 20,000. The LSTM model built is trained

with MSE loss (Equation (14) shows the MSE function). It creates a criterion whereby it measures the squared error (squared L2 norm) between each element in the input x and target y.

$$MSE = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n} \quad (14)$$

The uncertainty of parameters can be reduced with the use of robust optimization algorithms [78]. Optimization is critical in LSTM training as it involves finding the optimal set of parameters to minimize the loss function. One of the popular optimizers is the adaptive moment estimation (ADAM) optimizer which combines the benefits of the both momentum and adaptive learning rate methods. ADAM is a method for efficient stochastic optimization as it only requires first-order gradients. The amount of memory required is very little and it computes the individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients. ADAM has been shown to converge faster and achieve better performance than stochastic gradient descent (SGD) and its variants [79]. In the LSTM model that we built, the ADAM optimizer with an initial learning rate of 0.01 was applied.

A definite loop was implemented in Python to retrieve IMERG and rain gauge data from MariaDB, which is then fed into the LSTM modeling framework developed with Pytorch. We developed an automated program (refer to Section 2.3) to capture and record CC, $PBias$, MAE, $RMSE$, and KGE both before and after training the data with LSTM. These performance metrics were stored in a table in MariaDB.

2.3. Development of pHp Programs

In the present study, the pHp (recursive acronym for hypertext pre-processor) [80] is used to retrieve the data from MariaDB and display the results and graph plotting. The pHp is a widely used open source scripting language for web development and can be embedded into HTML [81]. Compared to Microsoft Excel, pHp is more dynamic and interactive. This study developed three pHp programs. The first pHp program generally presents the comparison of daily satellite estimations and rain gauge observations at the point-to-pixel scale. Figure 5 shows the interface of the program. To run the program, the user is required to select the grid/rain gauge point, key in, or select the starting and end dates for the program to query the result from the database. Once the criteria are selected, the program will display the result responsively based on the user's selected criteria. The comparison of the two datasets will be visualized in the form of a scatter plot and time series graph. The data points in both graphs represent the average rainfall in a grid. The statistical performance will also display at the bottom of the map.

Simple and straightforward step simulations are important in data modeling [82]. Maps generated for spatio-temporal analysis provide a clearer picture of the regions that required attention under current climate change [83]. The second pHp program to be developed allows the user to visualize the satellite estimations and rain gauge observations in the form of maps, as shown in Figure 6. In this program, the user will need to select the study period, then the rainfall will be plotted according to the satellite grid size. The legend will be dynamically generated. It makes the researcher easier to visualize the distribution of rainfall data from both rain gauge the observation and satellite estimations of the same period.

The last pHp program was developed to visualize the distribution of every statistic that was used to evaluate the performance of SPEs before and after LSTM deep learning improvement. This pHp program will be visualizing the performance metrics in the IMERG grids in maps by connecting to MariaDB where the results are being stored. The results from this program are presented in Section 3.

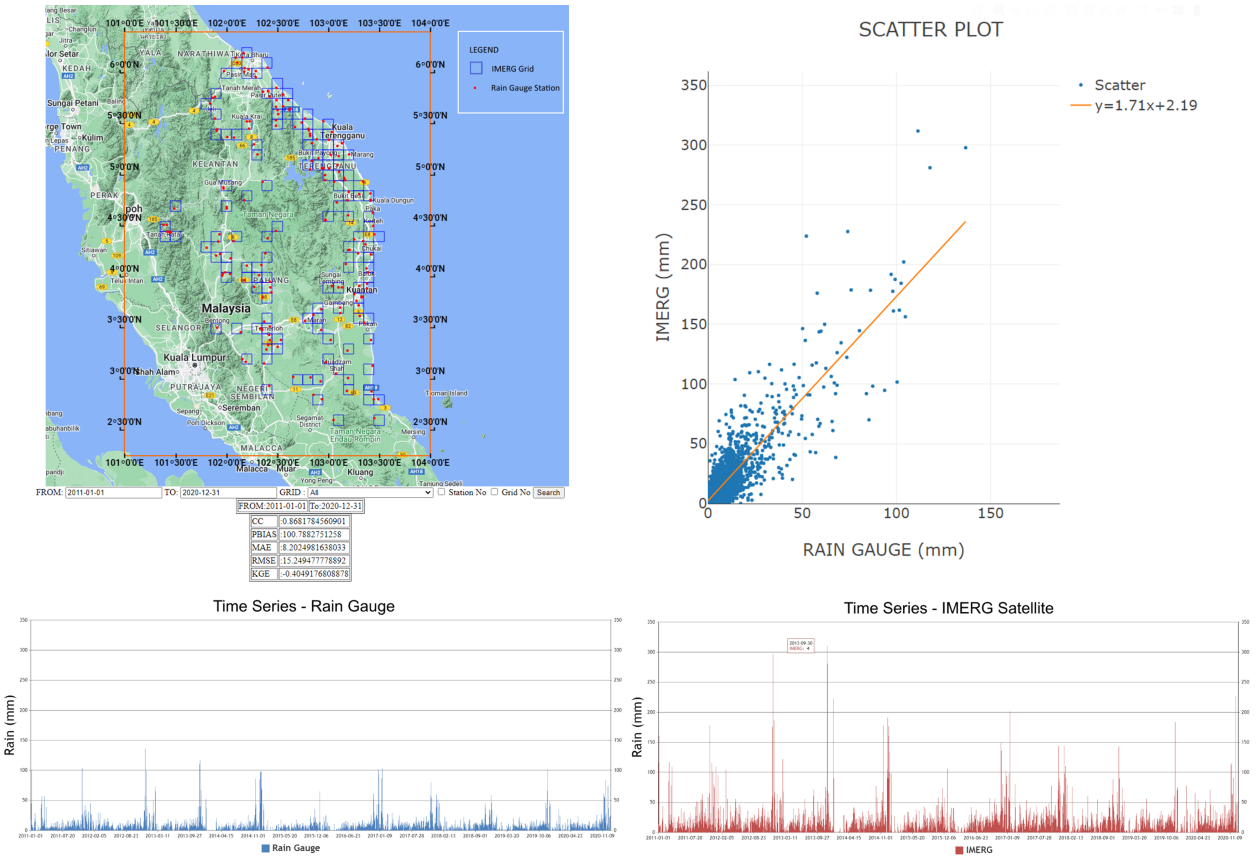

Figure 5. First pHp program that compares the daily satellite estimations and rain gauge observations at the point-to-pixel scale.

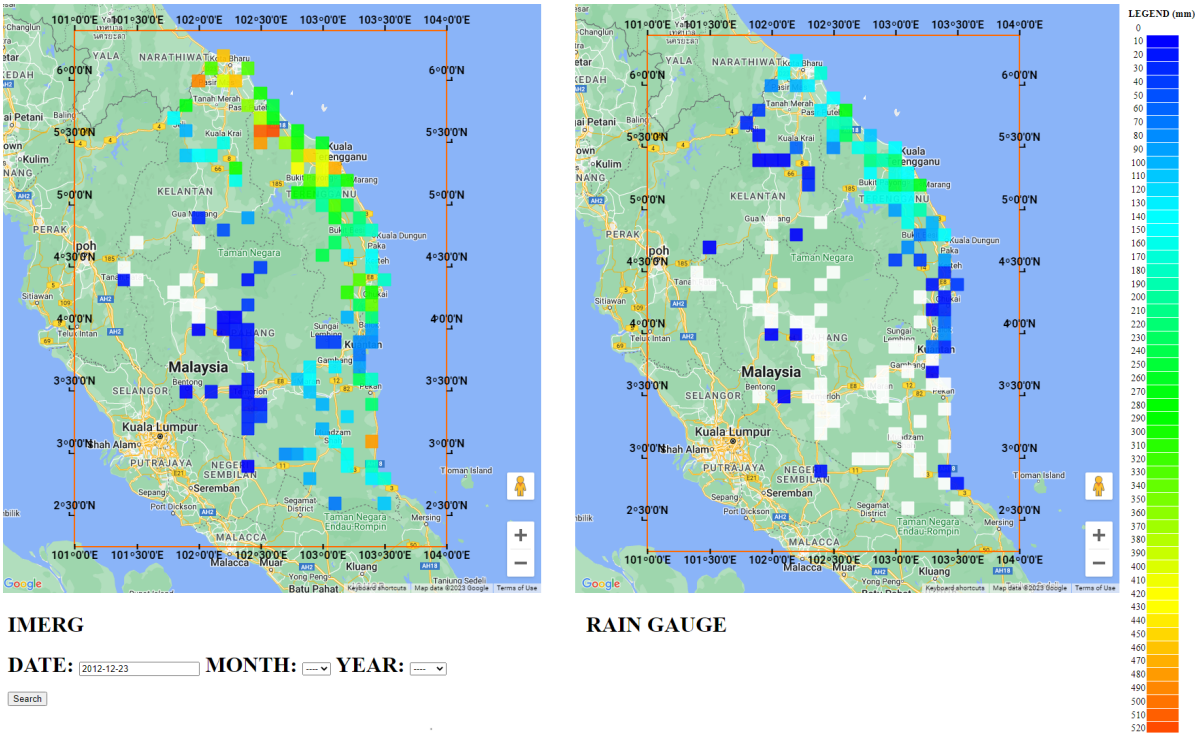

Figure 6. Second pHp program that shows the daily rainfall distribution of satellite estimations and rain gauge observations.

3. Results

By considering IMERG grids and time period selections in the first pHp program, researchers gain a comprehensive understanding of the factors influencing the observed results. From the terrain view, it is noticed that most of the rain gauges are positioned outside the mountainous areas, as shown in Figure 7. The terrain map clearly illustrates the spatial relationship between rain gauge locations and the mountainous terrain. This finding suggests that the rainfall measurements may not adequately capture the variability and intensity of the rainfall within the mountainous areas.

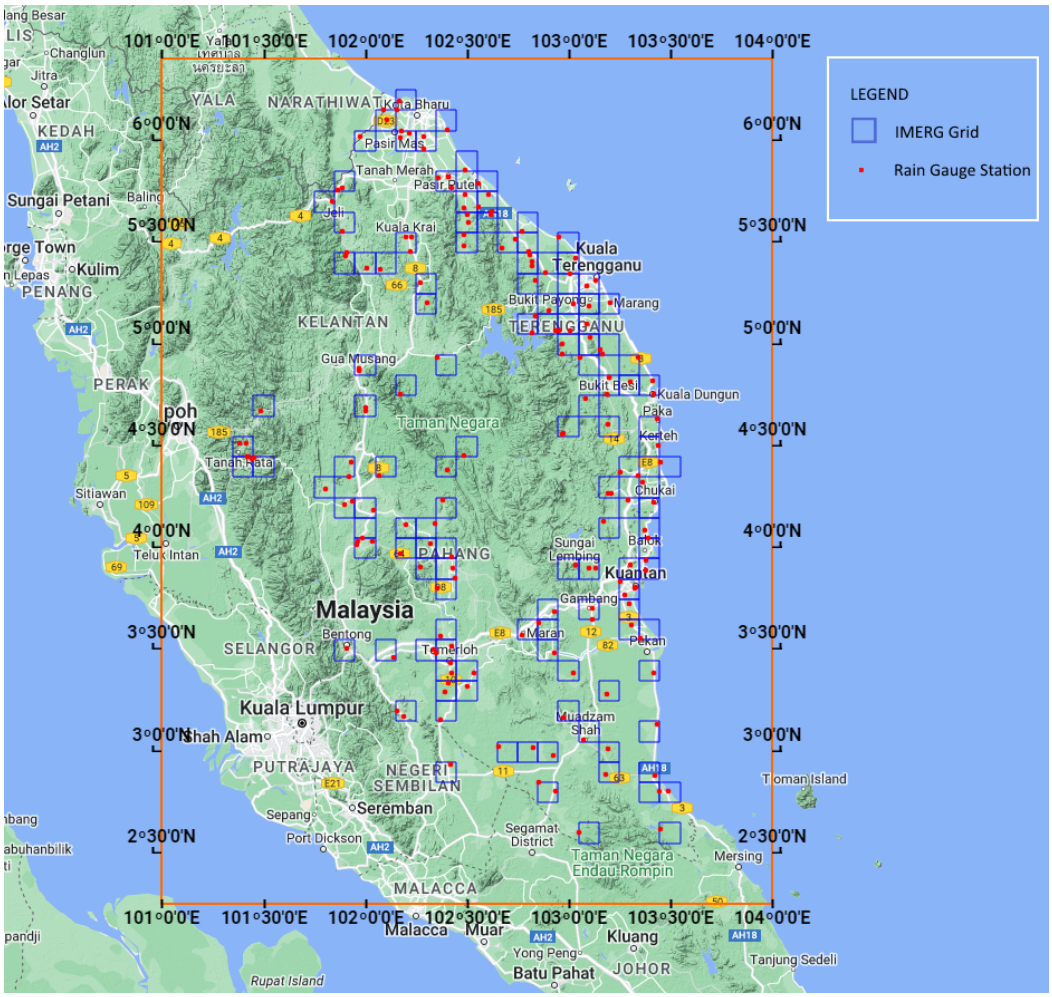

Figure 7. Selected IMERG grids that contain rain gauge stations.

Based on the regression result produced by the program, it is noted that the IMERG satellite rainfall estimations overestimated the actual rainfall by about 100% and have a high *RMSE* (15.25 mm/day). The analysis of the scatter plot reveals a noticeable relationship between IMERG overestimation with reference to the rain gauge data. The linear equation generated provides insights into the nature and magnitude of the overestimation. Referring to the scatter plot (Figure 8), the data points become more scattered as the rainfall intensity increases. In other words, the performance of the satellite estimations decreased as the rainfall became heavier.

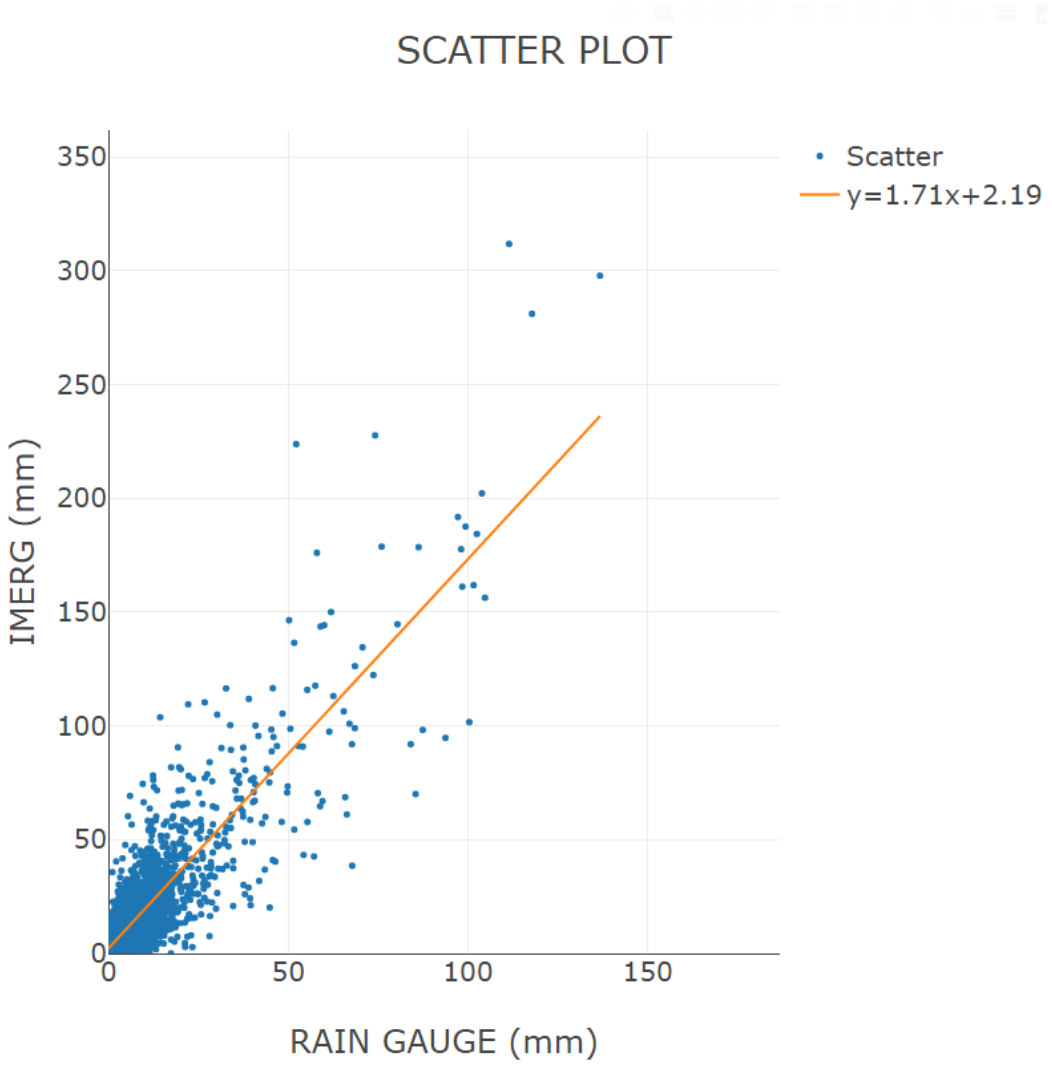

Figure 8. Scatter plot that shows the comparison of the daily areal rainfall between IMERG estimations and rain gauge observations for the period 2011–2020.

Referring to the time series graph generated (Figure 9), high peak rainfall was found in the month of December every year, which is also part of the Northeast monsoon. This is common as the prevailing north-easterly winds flow across the South China Sea, bringing in more moisture to the land surface. Soo, Jaafar [67] conducted a test to evaluate the capturing storm capabilities of satellites by increasing the rainfall threshold. They found that, regardless of any SPEs, their accuracy decreased as the rainfall threshold increased, which can be considered consistent with our present study.

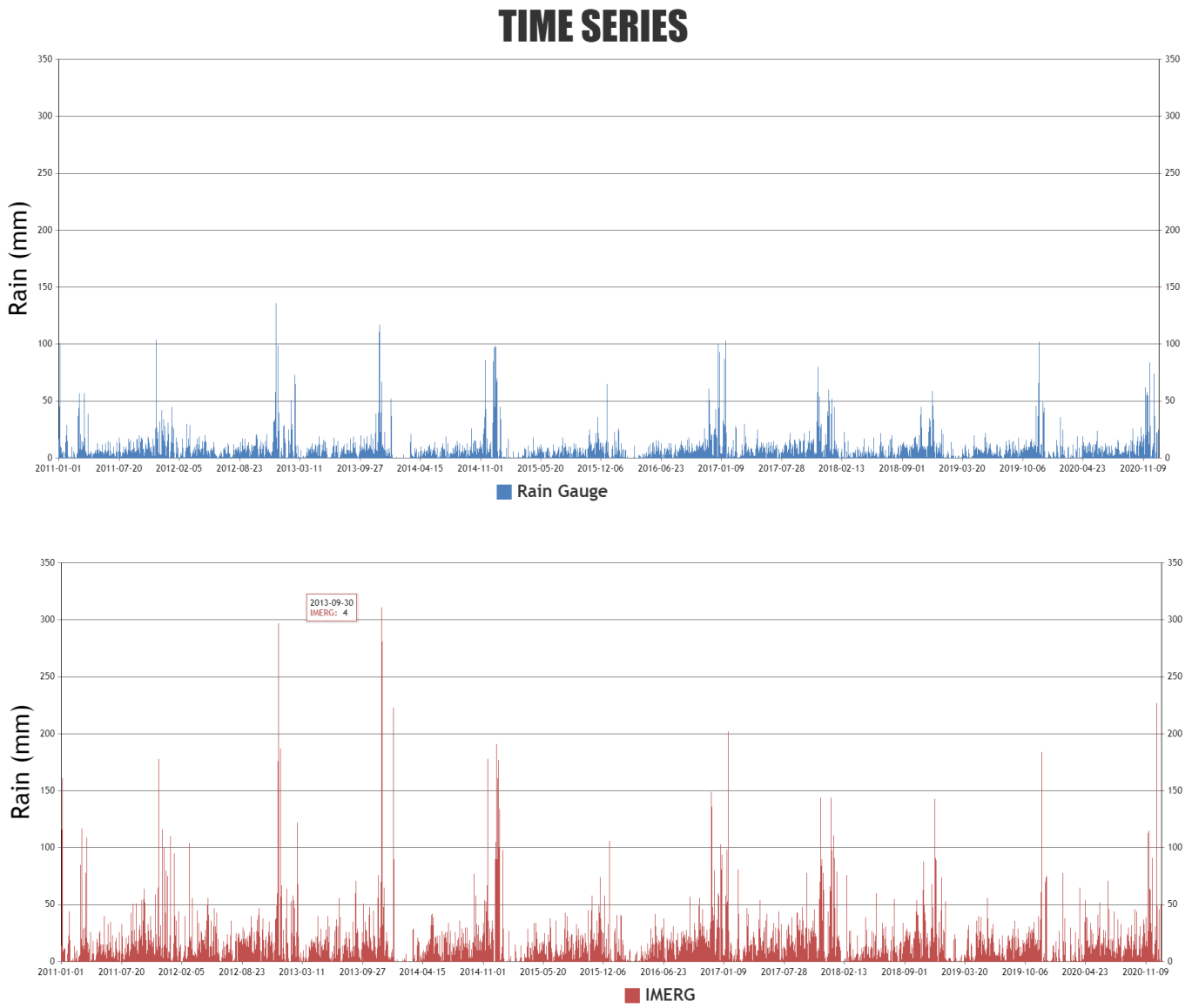

Figure 9. Time series graph for rain gauge observations and IMERG rainfall for the period 2011–2020.

From the second pHp program, it is noted that the hardest hit regions are along the northeast part of Peninsular Malaysia, as shown in Figure 10. Overall, IMERG satellite rainfall overestimates the actual rainfall for the whole study site.

The extracted rainfall estimations from the SPEs were then trained using the LSTM model programmed in Python with the rain gauge observations as the reference data. As a result, it was found that all of the statistics metrics improved after performing the LSTM training, indicating that the proposed LSTM methodology has good adaptability in enhancing the SPEs. The *CC* was raised by approximately 14.19%, *PBias* was reduced by 81–118%, *MAE* was reduced by 18–59%, *RMSE* was reduced by 1–66%, and the *KGE* was raised by approximately 200%. Table 1 summarizes the highest and lowest value of each metric for evaluating the performance of original and improved SPEs.

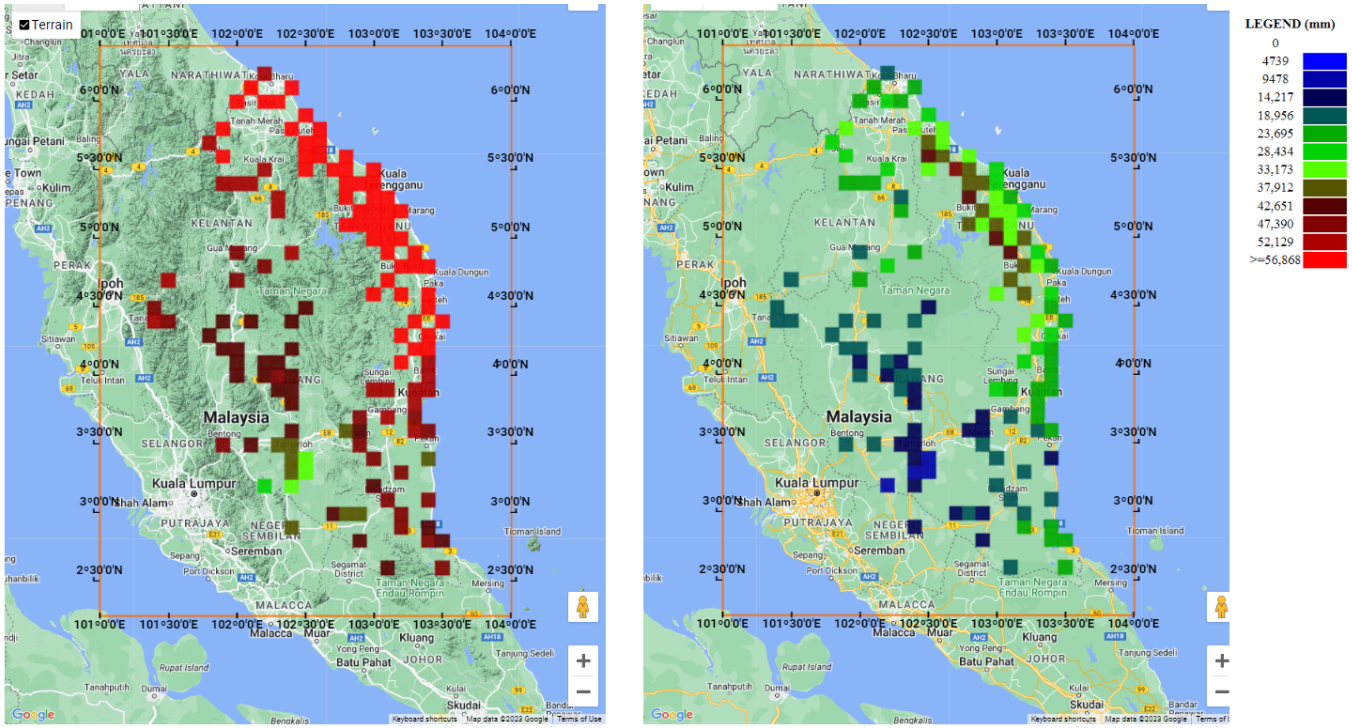

Figure 10. Total rainfall for IMERG estimations and rain gauge observations over each IMERG grid for the period 2011–2020.

Table 1. The highest and lowest performance of each statistic performance of SPEs for the period 2011–2020.

Statistic Metrics	Original SPEs (before Performing LSTM) Lowest	Original SPEs (before Performing LSTM) Highest	Enhanced SPEs (after Performing LSTM) Lowest	Enhanced SPEs (after Performing LSTM) Highest
CC	0.36	0.74	0.40	0.81
PBias (%)	38.05	229.50	−12.59	18.36
MAE	9.32	14.54	4.08	11.37
RMSE	19.40	33.85	8.11	33.51
KGE	−1.68	0.42	−0.35	0.77

The last pHp program, as shown in Figures 11–15, presents the performance metrics of all IMERG grids before and after LSTM training in the map format. Originally, the IMERG is able to capture the rainfall along the coastal region in the northeast part of Kelantan and Terengganu states with a *CC* ranging from 0.50 to 0.70, and poorer performance was found for the central region of Peninsular Malaysia (Pahang state) with a *CC* less than 0.50.

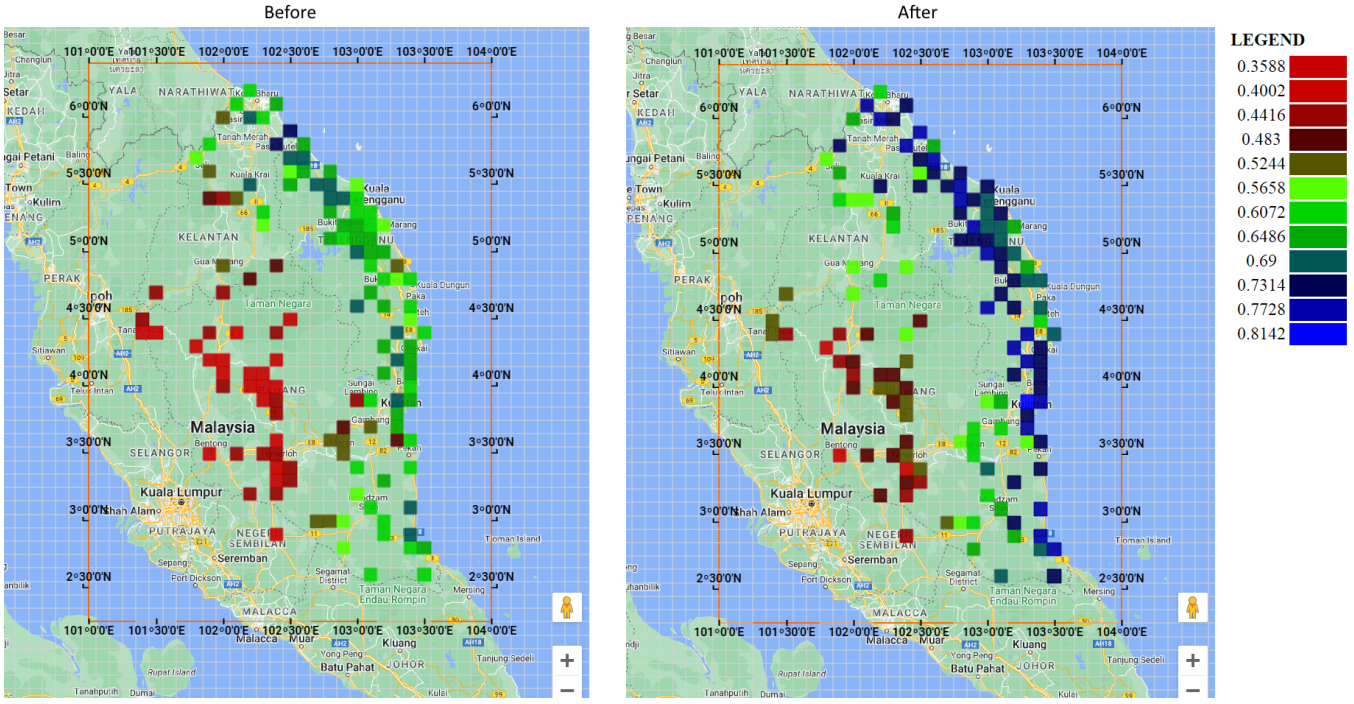

Figure 11. pHp program that shows the performance of *CC* before and after LSTM training for the period 2011–2020.

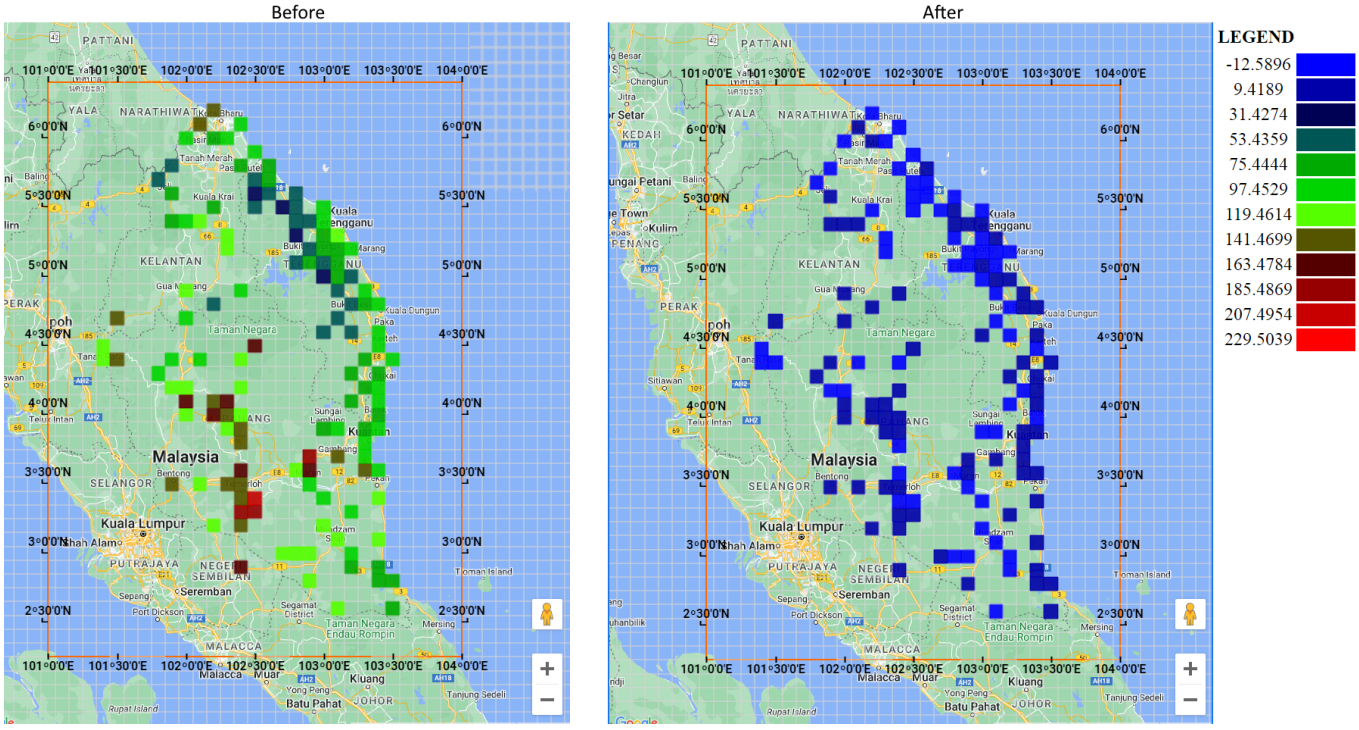

Figure 12. pHp program that shows the performance of *PBias* (%) before and after LSTM training for the period 2011–2020.

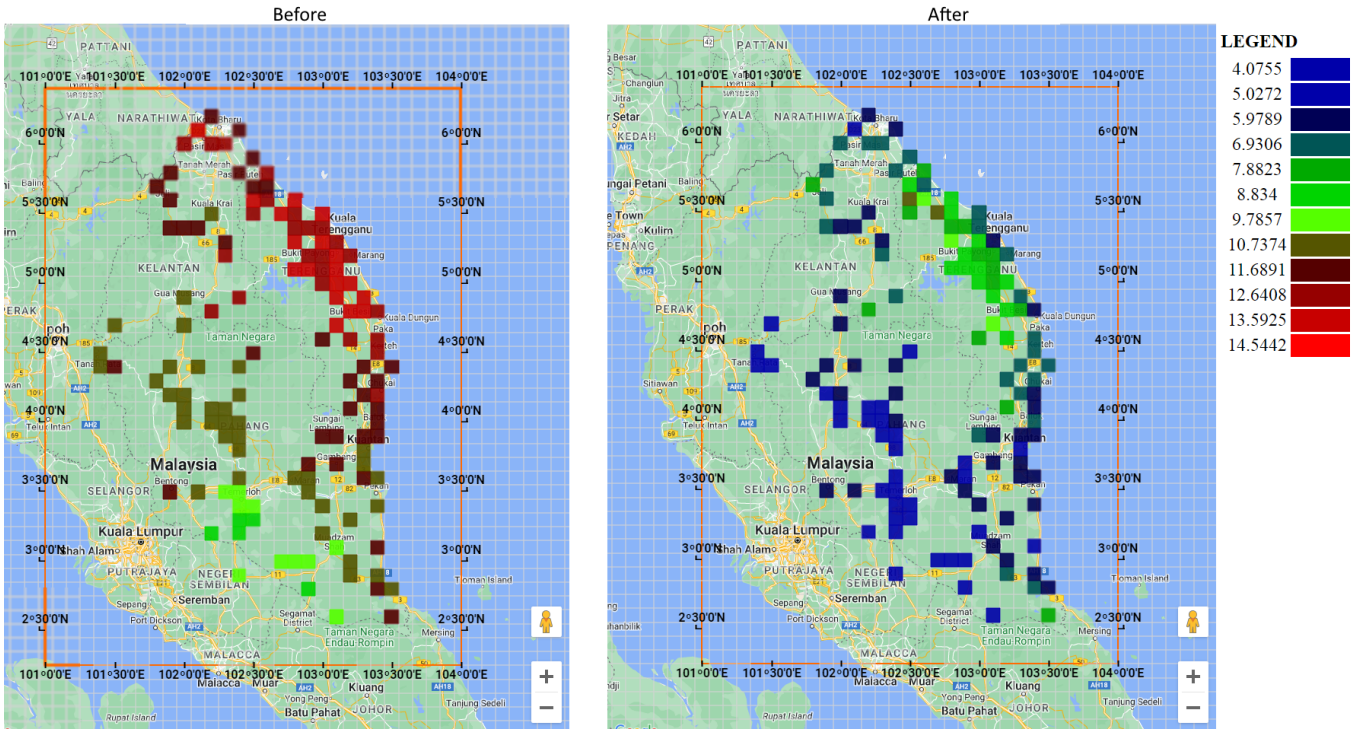

Figure 13. pHp program that shows the performance of *MAE* (mm/day) before and after LSTM training for the period 2011–2020.

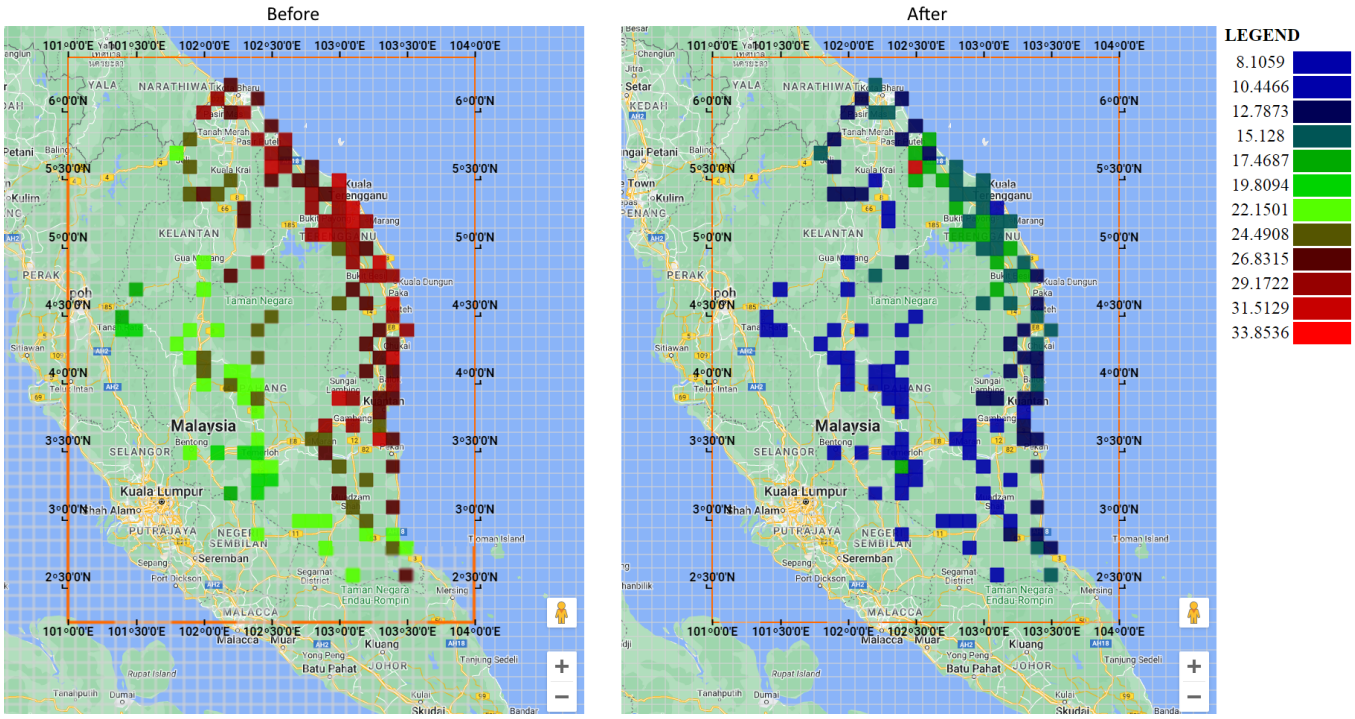

Figure 14. pHp program that shows the performance of *RMSE* (mm/day) before and after LSTM training for the period 2011–2020.

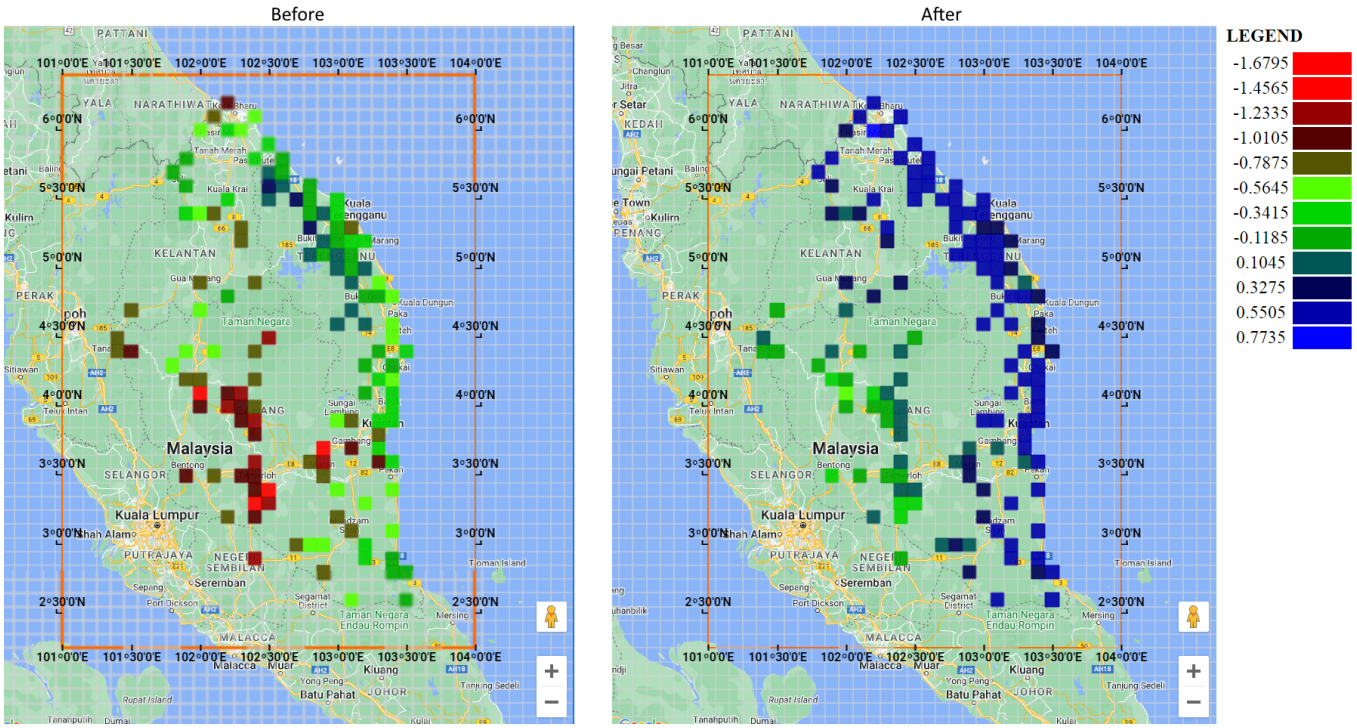

Figure 15. pHp program that shows the performance of *KGE* before and after LSTM training for the period 2011–2020.

4. Discussion

The management of satellite data poses significant challenges due to their voluminous nature, necessitating the implementation of efficient step-by-step protocols. Proper data storage is also crucial to accommodate large datasets. Streamlining data processing can optimize researcher time and productivity. Furthermore, visualizing the results may the enhance data analysis and facilitate informed decision making. In this study, we propose a sequential approach for data storage and processing utilizing Python, pHp, and MariaDB. Our methodology offers improved organization and structure to the data management including data processing, comparing and enhancing the satellite data with rain gauge observations. Apart from that, our results demonstrate the program's effectiveness in managing large datasets, thus providing a valuable tool for researchers dealing with data-intensive applications. In their recent study, Guo, B. et al. [84] evaluated the eight satellite precipitation products using similar performance metrics as those presented herein. The utilization of the proposed pHp programs could significantly enhance the visualization of the output.

The findings of this study highlight the potential of visualization to access the accuracy of IMERG estimations. This provides a framework for correcting and adjusting IMERG estimations based on the rain gauge measurements. Furthermore, the examination of different IMERG grid resolutions and time period provides a clearer picture of the impact of spatial-temporal factors on the observed relationships. The absence of rain gauges within mountainous areas may result in the misinterpretation of rainfall patterns. The spatial distribution of rain gauges outside the mountainous areas could impact the hydrological modeling, climate studies, and water resource management in mountainous areas. The reasons behind the uneven distribution of rain gauges in mountainous areas might be attributed to factors such as logistical challenges, the limitation of access, and high cost for installation and maintenance.

Although the LSTM model was able to enhance the SPEs, it is noted that the overall performance was not uniform. The non-uniform performance highlights the complexity of satellite precipitation estimates. Additionally, this study did not consider other geographical data as input such as climatic data such as wind speed, humidity, etc. Future research

should focus on identifying and addressing these factors to improve the uniformity of the performance. In terms of *CC* and *KGE*, those rainfall grids along the coastal region outperformed those in the central region. Conversely, the rainfall grids in the central region showed better performances in terms of *MAE* and *RMSE*. Perhaps a different model should be elaborated for those grids for these two regions with different training parameters, otherwise further study is required for performing rule-based optimization. Apart from that, the model for training the SPE data should introduce relevant auxiliary factors such as altitude, brightness, temperature, humidity, etc., so that the model would represent the exact condition of the study area, and a more accurate output can be obtained.

In the LSTM modeling program developed, the sliding windows derived from the time series data are used to perform short-term prediction with a fixed length of sequence. The power of LSTM lies in the recurrence of this process with the continuous sliding of the data windows to reduce the errors and bias between IMERG satellite rainfall estimates and rain-gauged observation data. With sufficient time series data (10 years of rainfall data), LSTM modeling with the ADAM optimizer has shown the great correction of IMERG satellite rainfall estimates that are trained with rain gauge observation data. Recent research performed to correct the bias of daily satellite precipitation estimates along the Kelantan River Basin using a deep learning approach managed to decrease the *RMSE* to a maximum of 30.0% and *MAE* to a maximum of 23.2%, respectively [85]. Although our current research area covers a much bigger area, we managed to reduce the *RMSE* and *MAE* to maxima of 66% and 59%, respectively. Further investigation can be conducted by using different optimizers to train the data with LSTM modeling. The machine learning algorithm processing time is crucial when dealing with huge datasets [86]. With the automated shell program developed, the model training job was scheduled and automatically run 24 by 7 and it has optimized the performance of the computing resource. Although the multiple layer implementation of the LSTM model is proven to provide better results [87], it could be challenging in making decisions regarding the model parameters during LSTM training. An existing dataset is stored in a relational database system and this will facilitate further research studies of extreme event forecasting with machine learning. Nevertheless, it is worthwhile to consider the potential benefits of utilizing non-relational database systems, such as MongoDB. The scalable data processing and analysis pipeline, especially wherein data are visualized in the pHp program and accessible via a relational database, will more effectively and efficiently facilitate the collaboration and data sharing among relevant researchers and data analysts online. Cloud-based platforms offer flexibility for deep learning architecture training [88]. By utilizing database connectors and the application programming interface (API), AI analysis tools can establish connections and conduct comprehensive analysis as well as forecasting modeling. The development of reliable flood forecasting models can be very helpful in data-scarce regions [89].

5. Conclusions

This study created a sequential way of systematically and efficiently managing the huge amount of data retrieved from SPEs and rain gauge observations. This study investigated the optimal way to effectively integrate the separate inputs and merge them into a single dataset. This could potentially be changing the way researchers conduct their conventional research. The motivation behind this method is to enable researchers to quickly process and visualize large datasets to better analyze the dataset by using a sequential of programs such as Python, pHp, SQL, etc. Rainfall prediction can be easily explored and trained with these programs. In the present study, the LSTM is also proposed in enhancing the accuracy of the SPEs. This model is a great neural network model that has the advantage of processing data with time series elements. Based on the outputs, the LSTM with the ADAM optimizer showed the capability of improving the accuracy of SPEs. The variations observed across different grids and time periods provides valuable ground for further investigation into the factors influencing the effectiveness of the model. As the degree of improvement was not at the same level for the whole study area, it is

recommended to perform a different model for those regions with a poorer performance with the different auxiliary factors included. Deep learning model training for rainfall data requires a systematic methodology as it is very time-consuming and could be experimentally oriented. Different comparative data scales such as weekly and monthly precipitation scores can be the consideration for future studies using different machine learning approaches. This comprehensive and scalable data processing and analysis pipeline used for deep learning modeling can be applicable for other areas such as flood forecasting and modeling, agricultural harvesting forecast, soil erosion, and land slide prediction in any region that have such domain challenges. Future research should also consider the use of LSTM regularization techniques with the identification of different penalty terms for the loss function with optimized hyperparameter sets. Dropout and early stopping could prevent overfitting and improve the generalization ability of LSTM models.

Author Contributions: Conceptualization, S.H.L., M.M., F.Y.T., E.Z.X.S. and S.C.T.; methodology, S.C.T., S.H.L. and M.M.; software, S.C.T.; Validation, S.H.L. and M.M.; formal analysis, E.Z.X.S. and S.C.T.; investigation, F.Y.T.; resources, S.H.L.; data curation, M.M.; writing—original draft preparation, S.C.T.; writing—review and editing, E.Z.X.S., S.H.L., M.M. and F.Y.T.; visualization, S.C.T.; supervision, S.H.L., M.M. and F.Y.T.; project administration, S.C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study are available from the corresponding authors upon reasonable request.

Acknowledgments: The authors would like to thank the Malaysian Department of Irrigation and Drainage (DID) for providing the daily precipitation data as well as the developers of the satellite estimations for providing the downloadable data.

Conflicts of Interest: The authors declare they have no competing interest.

Abbreviations

The following abbreviations are used in this manuscript:

IMERG	Integrated Multi-Satellite Retrievals for GPM
LSTM	Long Short-Term Memory
ADAM	Adaptive Moment Estimation
MAE	Mean Square Error
RMSE	Root-Mean-Square Error
KGE	Kling–Gupta Efficiency
SQL	Structured Query Language
pHp	Hypertext Pre-processor
ML	Machine Learning
DL	Deep Learning
SPE	Satellite Precipitation Estimation
RNN	Recurrent Neural Network
CSV	Comma Separated Values
DID	Department of Irrigation and Drainage
EXE	Executable
HTML	Hypertext Markup Language

References

1. Eltahir, E.A.B.; Bras, R.L. Precipitation recycling. *Rev. Geophys.* **1996**, *34*, 367–378. [CrossRef]
2. Kidd, C. Satellite rainfall climatology: A review. *Int. J. Climatol.* **2001**, *21*, 1041–1066. [CrossRef]
3. Kidd, C.; Becker, A.; Huffman, G.J.; Muller, C.L.; Joe, P.; Skofronick-Jackson, G.; Kirschbaum, D.B. So, How Much of the Earth's Surface Is Covered by Rain Gauges? *Bull. Am. Meteorol. Soc.* **2017**, *98*, 69–78. [CrossRef]
4. Ma, Q.; Li, Y.; Feng, H.; Yu, Q.; Zou, Y.; Liu, F.; Pulatov, B. Performance evaluation and correction of precipitation data using the 20-year IMERG and TMPA precipitation products in diverse subregions of China. *Atmos. Res.* **2021**, *249*, 105304. [CrossRef]
5. Jiang, S.; Ren, L.; Hong, Y.; Yong, B.; Yang, X.; Yuan, F.; Ma, M. Comprehensive evaluation of multi-satellite precipitation products with a dense rain gauge network and optimally merging their simulated hydrological flows using the Bayesian model averaging method. *J. Hydrol.* **2012**, *452–453*, 213–225. [CrossRef]
6. Gebremicael, T.G.; Mohamed, Y.A.; Zaag, P.v.; Berhe, A.G.; Haile, G.G.; Hagos, E.Y.; Hagos, M.K. Comparison and validation of eight satellite rainfall products over the rugged topography of Tekeze-Atbara Basin at different spatial and temporal scales. *Hydrol. Earth Syst. Sci. Discuss.* **2017**, *2017*, 1–31.
7. Turini, N.; Thies, B.; Bendix, J. Estimating High Spatio-Temporal Resolution Rainfall from MSG1 and GPM IMERG Based on Machine Learning: Case Study of Iran. *Remote Sens.* **2019**, *11*, 2307. [CrossRef]
8. Hong, Y.; Zhang, Y.; Khan, S. *Hydrologic Remote Sensing: Capacity Building for Sustainability and Resilience*; CRC Press: Boca Raton, FL, USA, 2016.
9. Llauca, H.; Lavado-Casimiro, W.; León, K.; Jimenez, J.; Traverso, K.; Rau, P. Assessing Near Real-Time Satellite Precipitation Products for Flood Simulations at Sub-Daily Scales in a Sparsely Gauged Watershed in Peruvian Andes. *Remote Sens.* **2021**, *13*, 826. [CrossRef]
10. Kim, T.; Yang, T.; Zhang, L.; Hong, Y. Near real-time hurricane rainfall forecasting using convolutional neural network models with Integrated Multi-satellitE Retrievals for GPM (IMERG) product. *Atmos. Res.* **2022**, *270*, 106037. [CrossRef]
11. Liu, C.-Y.; Aryastana, P.; Liu, G.-R.; Huang, W.-R. Assessment of satellite precipitation product estimates over Bali Island. *Atmos. Res.* **2020**, *244*, 105032. [CrossRef]
12. Huffman, G.J.; Bolvin, D.T.; Nelkin, E.J.; Wolff, D.B.; Adler, R.F.; Gu, G.; Hong, Y.; Bowman, K.P.; Stocker, E.F. The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.* **2007**, *8*, 38–55. [CrossRef]
13. Joyce, R.J.; Janowiak, J.E.; Arkin, P.A.; Xie, P. CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeorol.* **2004**, *5*, 487–503. [CrossRef]
14. Hsu, K.-l., Gao, X.; Sorooshian, S.; Gupta, H.V. Precipitation estimation from remotely sensed information using artificial neural networks. *J. Appl. Meteorol.* **1997**, *36*, 1176–1190. [CrossRef]
15. Sorooshian, S.; Hsu, K.-L.; Gao, X.; Gupta, H.V.; Imam, B.; Braithwaite, D. Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bull. Am. Meteorol. Soc.* **2000**, *81*, 2035–2046. [CrossRef]
16. Kubota, T.; Shige, S.; Hashizume, H.; Aonashi, K.; Takahashi, N.; Seto, S.; Hirose, M.; Takayabu, Y.N.; Ushio, T.; Nakagawa, K. Global precipitation map using satellite-borne microwave radiometers by the GSMaP project: Production and validation. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 2259–2275. [CrossRef]
17. Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A. The climate hazards infrared precipitation with stations—A new environmental record for monitoring extremes. *Sci. Data* **2015**, *2*, 1–21. [CrossRef] [PubMed]
18. Huffman, G.J.; Bolvin, D.T.; Braithwaite, D.; Hsu, K.-L.; Joyce, R.J.; Kidd, C.; Nelkin, E.J.; Sorooshian, S.; Stocker, E.F.; Tan, J. Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG). In *Satellite Precipitation Measurement*; Springer: Cham, Switzerland, 2020; pp. 343–353.
19. Ma, Y.; Sun, X.; Chen, H.; Hong, Y.; Zhang, Y. A two-stage blending approach for merging multiple satellite precipitation estimates and rain gauge observations: An experiment in the northeastern Tibetan Plateau. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 359–374. [CrossRef]
20. Arshad, M.; Ma, X.; Yin, J.; Ullah, W.; Ali, G.; Ullah, S.; Liu, M.; Shahzaman, M.; Ullah, I. Evaluation of GPM-IMERG and TRMM-3B42 precipitation products over Pakistan. *Atmos. Res.* **2021**, *249*, 105341. [CrossRef]
21. Moazami, S. and Najafi, M.R. A comprehensive evaluation of GPM-IMERG V06 and MRMS with hourly ground-based precipitation observations across Canada. *J. Hydrol.* **2021**, *594*, 125929. [CrossRef]
22. Mekonnen, K.; Melesse, A.M.; Woldesenbet, T.A. Spatial evaluation of satellite-retrieved extreme rainfall rates in the Upper Awash River Basin, Ethiopia. *Atmos. Res.* **2021**, *249*, 105297. [CrossRef]
23. Huang, W.-R.; Liu, P.-Y.; Hsu, J.; Li, X.; Deng, L. Assessment of Near-Real-Time Satellite Precipitation Products from GSMaP in Monitoring Rainfall Variations over Taiwan. *Remote Sens.* **2021**, *13*, 202. [CrossRef]
24. Gan, F.; Gao, Y.; Xiao, L. Comprehensive validation of the latest IMERG V06 precipitation estimates over a basin coupled with coastal locations, tropical climate and hill-karst combined landform. *Atmos. Res.* **2021**, *249*, 105293. [CrossRef]
25. Nepal, B.; Shrestha, D.; Sharma, S.; Shrestha, M.S.; Aryal, D.; Shrestha, N. Assessment of GPM-Era Satellite Products' (IMERG and GSMaP) Ability to Detect Precipitation Extremes over Mountainous Country Nepal. *Atmosphere* **2021**, *12*, 254. [CrossRef]
26. Palpanas, T. Data Series Management: The Road to Big Sequence Analytics. *SIGMOD Rec.* **2015**, *44*, 47–52. [CrossRef]

27. Huntington, J.L.; Hegewisch, K.C.; Daudert, B.; Morton, C.G.; Abatzoglou, J.T.; McEvoy, D.J.; Erickson, T. Climate Engine: Cloud Computing and Visualization of Climate and Remote Sensing Data for Advanced Natural Resource Monitoring and Process Understanding. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 2397–2410. [CrossRef]
28. Yin, J.; Guo, S.; Gu, L.; Zeng, Z.; Liu, D.; Chen, J.; Shen, Y.; Xu, C.-Y. Blending multi-satellite, atmospheric reanalysis and gauge precipitation products to facilitate hydrological modelling. *J. Hydrol.* **2021**, *593*, 125878. [CrossRef]
29. Amani, M.; Ghorbanian, A.; Ahmadi, S.A.; Kakooei, M.; Moghimi, A.; Mirmazloumi, S.M.; Moghaddam, S.H.A.; Mahdavi, S.; Ghahremanloo, M.; Parsian, S.; et al. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5326–5350. [CrossRef]
30. Chen, H.; Chandrasekar, V.; Cifelli, R.; Xie, P. A Machine Learning System for Precipitation Estimation Using Satellite and Ground Radar Network Observations. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 982–994. [CrossRef]
31. Zhang, L.; Li, X.; Zheng, D.; Zhang, K.; Ma, Q.; Zhao, Y.; Ge, Y. Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. *J. Hydrol.* **2021**, *594*, 125969. [CrossRef]
32. Sadeghi, M.; Nguyen, P.; Hsu, K.; Sorooshian, S. Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. *Environ. Model. Softw.* **2020**, *134*, 104856. [CrossRef]
33. Kühnlein, M.; Appelhans, T.; Thies, B.; Nauss, T. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI. *Remote Sens. Environ.* **2014**, *141*, 129–143. [CrossRef]
34. Srivastava, S. and Lessmann, S. A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data. *Sol. Energy* **2018**, *162*, 232–247. [CrossRef]
35. Graves, A.; Jaitly, N.; Mohamed, A.R. Hybrid speech recognition with deep bidirectional LSTM. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013.
36. Wang, S.; Jiang, J. Learning natural language inference with LSTM. *arXiv* **2015**, arXiv:1512.08849.
37. Karevan, Z.; Suykens, J.A.Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Netw.* **2020**, *125*, 1–9. [CrossRef] [PubMed]
38. DiPietro, R.; Hager, G.D. Chapter 21—Deep learning: RNNs and LSTM. In *Handbook of Medical Image Computing and Computer Assisted Intervention*; Zhou, S.K., Rueckert, D., Fichtinger, G., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 503–519.
39. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
40. Akbari Asanjan, A.; Yang, T.; Hsu, K.; Sorooshian, S.; Lin, J.; Peng, Q. Short-Term Precipitation Forecast Based on the PERSIANN System and LSTM Recurrent Neural Networks. *J. Geophys. Res. Atmos.* **2018**, *123*, 12543–12563. [CrossRef]
41. Miao, Q.; Pan, B.; Wang, H.; Hsu, K.; Sorooshian, S. Improving Monsoon Precipitation Prediction Using Combined Convolutional and Long Short Term Memory Neural Network. *Water* **2019**, *11*, 977. [CrossRef]
42. Wu, H.; Yang, Q.; Liu, J.; Wang, G. A spatiotemporal deep fusion model for merging satellite and gauge precipitation in China. *J. Hydrol.* **2020**, *584*, 124664. [CrossRef]
43. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
44. Dehghani, A.; Moazam, H.M.Z.H.; Mortazavizadeh, F.; Ranjbar, V.; Mirzaei, M.; Mortezavi, S.; Ng, J.L.; Dehghani, A. Comparative evaluation of LSTM, CNN, and ConvLSTM for hourly short-term streamflow forecasting using deep learning approaches. *Ecol. Inform.* **2023**, *75*, 102119. [CrossRef]
45. Sharma, O. Deep challenges associated with deep learning. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 14–16 February 2019.
46. Gao, W.; Gao, J.; Yang, L.; Wang, M.; Yao, W. A novel modeling strategy of weighted mean temperature in China using RNN and LSTM. *Remote Sens.* **2021**, *13*, 3004. [CrossRef]
47. Gers, F.A.; Schraudolph, N.N.; Schmidhuber, J. Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* **2002**, *3*, 115–143.
48. Noh, S.-H. Analysis of gradient vanishing of RNNs and performance comparison. *Information* **2021**, *12*, 442. [CrossRef]
49. Mirzaei, M.; Yu, H.; Dehghani, A.; Galavi, H.; Shokri, V.; Karimi, S.M.; Sookhak, M. A novel stacked long short-term memory approach of deep learning for streamflow simulation. *Sustainability* **2021**, *13*, 13384. [CrossRef]
50. Mohsenzadeh Karimi, S.; Mirzaei, M.; Dehghani, A.; Galavi, H.; Huang, Y.F. Hybrids of machine learning techniques and wavelet regression for estimation of daily solar radiation. *Stoch. Environ. Res. Risk Assess.* **2022**, *36*, 4255–4269. [CrossRef]
51. Roh, Y.; Heo, G.; Whang, S.E. A Survey on Data Collection for Machine Learning: A Big Data—AI Integration Perspective. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 1328–1347. [CrossRef]
52. Woldemeskel, F.M.; Sivakumar, B.; Sharma, A. Merging gauge and satellite rainfall with specification of associated uncertainty across Australia. *J. Hydrol.* **2013**, *499*, 167–176. [CrossRef]
53. Villalobos-Herrera, R.; Blenkinsop, S.; Guerreiro, S.B.; O'Hara, T.; Fowler, H.J. Sub-hourly resolution quality control of rain-gauge data significantly improves regional sub-daily return level estimates. *Q. J. R. Meteorol. Soc.* **2022**, *148*, 3252–3271. [CrossRef] [PubMed]
54. Hsu, J.; Huang, W.-R.; Liu, P.-Y.; Li, X. Validation of CHIRPS Precipitation Estimates over Taiwan at Multiple Timescales. *Remote Sens.* **2021**, *13*, 254. [CrossRef]
55. Bhuiyan, M.A.E.; Yang, F.; Biswas, N.K.; Rahat, S.H.; Neelam, T.J. Machine learning-based error modeling to improve GPM IMERG precipitation product over the brahmaputra river basin. *Forecasting* **2020**, *2*, 248–266. [CrossRef]

56. Lazri, M.; Labadi, K.; Brucker, J.M.; Ameur, S. Improving satellite rainfall estimation from MSG data in Northern Algeria by using a multi-classifier model based on machine learning. *J. Hydrol.* **2020**, *584*, 124705. [CrossRef]
57. Mayowa, O.O.; Pour, S.H.; Shahid, S.; Mohsenipour, M.; Harun, S.B.; Heryansyah, A.; Ismail, T. Trends in rainfall and rainfall-related extremes in the east coast of peninsular Malaysia. *J. Earth Syst. Sci.* **2015**, *124*, 1609–1622. [CrossRef]
58. Juneng, L.; Tangang, F.; Reason, C. Numerical case study of an extreme rainfall event during 9–11 December 2004 over the east coast of Peninsular Malaysia. *Meteorol. Atmos. Phys.* **2007**, *98*, 81–98. [CrossRef]
59. Hai, O.S.; Samah, A.A.; Chenoli, S.N.; Subramaniam, K.; Mazuki, M.Y.A. Extreme rainstorms that caused devastating flooding across the east coast of Peninsular Malaysia during November and December 2014. *Weather. Forecast.* **2017**, *32*, 849–872. [CrossRef]
60. Svennerberg, G. *Beginning Google Maps API 3*; Apress: New York, NY, USA, 2010.
61. Hu, L.; He, Z.; Liu, J.; Zheng, C. Method for Measuring the Information Content of Terrain from Digital Elevation Models. *Entropy* **2015**, *17*, 7021–7051. [CrossRef]
62. Collette, A. *Python and HDF5: Unlocking Scientific Data*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2013.
63. Van Rossum, G.; Drake, F. *Python 3 Reference Manual Createspace*; CreateSpace: Scotts Valley, CA, USA, 2009.
64. Witham, M.; Bender, I.; Gomes, R. Comparative Analysis of MariaDB's Performance Efficiency as a Suitable Replacement for MySQL. In Proceedings of the 2019 Midwest Instruction and Computing Symposiu, Fargo, ND, USA, 5–6 April 2019.
65. Lindstrom, J.; Das, D.; Mathiasen, T.; Arteaga, D.; Talagala, N. NVM aware MariaDB database system. In Proceedings of the 2015 IEEE Non-Volatile Memory System and Applications Symposium (NVMSA), Hong Kong, China, 19–21 August 2015.
66. Jamison, D.C. Structured Query Language (SQL) Fundamentals. *Curr. Protoc. Bioinform.* **2003**, 9.2.1–9.2.29. [CrossRef] [PubMed]
67. Soo, E.Z.X.; Jaafar, W.Z.W.; Lai, S.H.; Islam, T.; Srivastava, P. Evaluation of satellite precipitation products for extreme flood events: Case study in Peninsular Malaysia. *J. Water Clim. Chang.* **2018**, *10*, 871–892. [CrossRef]
68. Bathelemy, R.; Brigode, P.; Boisson, D.; Tric, E. Rainfall in the Greater and Lesser Antilles: Performance of five gridded datasets on a daily timescale. *J. Hydrol. Reg. Stud.* **2022**, *43*, 101203. [CrossRef]
69. Lee Rodgers, J.; Nicewander, W.A. Thirteen Ways to Look at the Correlation Coefficient. *Am. Stat.* **1988**, *42*, 59–66. [CrossRef]
70. Yassin, F.; Razavi, S.; Wheater, H.; Sapriza-Azuri, G.; Davison, B.; Pietroniro, A. Enhanced identification of a hydrologic model using streamflow and satellite water storage data: A multicriteria sensitivity analysis and optimization approach. *Hydrol. Processes* **2017**, *31*, 3320–3333. [CrossRef]
71. Strauch, M.; Kumar, R.; Eisner, S.; Mulligan, M.; Reinhardt, J.; Santini, W.; Vetter, T.; Friesen, J. Adjustment of global precipitation data for enhanced hydrologic modeling of tropical Andean watersheds. *Clim. Chang.* **2017**, *141*, 547–560. [CrossRef]
72. Pontius, R.G.; Thontteh, O.; Chen, H. Components of information for multiple resolution comparison between maps that share a real variable. *Environ. Ecol. Stat.* **2008**, *15*, 111–142. [CrossRef]
73. Willmott, C.J.; Matsuura, K. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 89–102. [CrossRef]
74. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
75. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [CrossRef]
76. Kling, H.; Fuchs, M.; Paulin, M. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J. Hydrol.* **2012**, *424–425*, 264–277. [CrossRef]
77. Sak, H.; Senior, A.; Beaufays, F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv* **2014**, arXiv:1402.1128.
78. Bazrafshan, O.; Ehteram, M.; Latif, S.D.; Huang, Y.F.; Teo, F.Y.; Ahmed, A.N.; El-Shafie, A. Predicting crop yields using a new robust Bayesian averaging model based on multiple hybrid ANFIS and MLP models. *Ain Shams Eng. J.* **2022**, *13*, 101724. [CrossRef]
79. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
80. Bakken, S.S.; Suraski, Z.; Schmid, E. *PHP Manual: Volume 2*; IUniverse, Incorporated: Bloomington, Indiana, 2000.
81. Ahmad, D.K.; Ahmad, M.F.; Ahmad, M.N.; Ahmad, A.S. An Experiment of Animation Development in Hypertext Preprocessor (PHP) and Hypertext Markup Language (HTML). *Int. J. Sci. Res. Comput. Sci. Eng.* **2020**, *8*, 45–51.
82. Lv, X.; Liu, B.; Yuan, D.; Feng, H.; Teo, F.Y. Random walk method for modeling water exchange: An application to coastal zone environmental management. *J. Hydro-Environ. Res.* **2016**, *13*, 66–75. [CrossRef]
83. Fung, K.F.; Chew, K.S.; Huang, Y.F.; Ahmed, A.N.; Teo, F.Y.; Ng, J.L.; Elshafie, A. Evaluation of spatial interpolation methods and spatiotemporal modeling of rainfall distribution in Peninsular Malaysia. *Ain Shams Eng. J.* **2022**, *13*, 101571. [CrossRef]
84. Guo, B.; Xu, T.; Yang, Q.; Zhang, J.; Dai, Z.; Deng, Y.; Zou, J. Multiple Spatial and Temporal Scales Evaluation of Eight Satellite Precipitation Products in a Mountainous Catchment of South China. *Remote Sens.* **2023**, *15*, 1373. [CrossRef]
85. Yang, X.; Yang, S.; Tan, M.L.; Pan, H.; Zhang, H.; Wang, G.; He, R.; Wang, Z. Correcting the bias of daily satellite precipitation estimates in tropical regions using deep neural network. *J. Hydrol.* **2022**, *608*, 127656. [CrossRef]
86. Meyer, H.; Kühnlein, M.; Appelhans, T.; Nauss, T. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. *Atmos. Res.* **2016**, *169*, 424–433. [CrossRef]
87. Gamboa-Villafruela, C.J.; Fernández-Alvarez, J.C.; Márquez-Mijares, M.; Pérez-Alarcón, A.; Batista-Leyva, A.J. Convolutional lstm architecture for precipitation nowcasting using satellite data. *Environ. Sci. Proc.* **2021**, *8*, 33.

88. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 1–74. [CrossRef] [PubMed]
89. Yeditha, P.K.; Kasi, V.; Rathinasamy, M.; Agarwal, A. Forecasting of extreme flood events using different satellite precipitation products and wavelet-based machine learning methods. *Chaos Interdiscip. J. Nonlinear Sci.* **2020**, *30*, 063115. [CrossRef] [PubMed]