An Assimilating Model Using Broad Learning System for Incorporating Multi-Source Precipitation Data With Environmental Factors Over Southeast China

Yuanyuan Zhou¹, Xu Li², Qiuhong Tang³,⁴, Sin Chi Kuok¹, Kai Fei¹, and Liang Gao¹

¹State Key Laboratory of Internet of Things for Smart City and Department of Civil and Environmental Engineering, University of Macau, Macao, China, ²School of Civil Engineering, Beijing Jiaotong University, Beijing, China, ³Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, ⁴University of Chinese Academy of Sciences, Beijing, China

Abstract Remote sensing technique is beneficial for rainfall data retrievals, however, enhancing the accuracy remains a challenge. In this study, a novel framework based on a broad learning system (BLS) was proposed to assimilate multi-source data. The dataset includes six satellite-based rainfall products (3B42V7, 3B42RT, IMERG, CBLD, GSMaP, and PCDR), gauge-based rainfall, and environmental data (temperature, specific humidity, wind speed, and locations) from 1 March 2014 to 31 December 2017 over southeast China (SEC). Leave-one-year-out cross-validation (LOYOCV) and independent validation were used to evaluate the BLS assimilating model. The proposed BLS model outperformed six original satellite-based products on Pearson's correlation coefficient (CC), root-mean-square error (RMSE), and Nash-Sutcliffe coefficient of efficiency (NSE) in each test year of LOYOCV. BLS model considering the environmental factors performed better on CC, RMSE, and NSE compared to that without environmental factors. Seasonal variations of daily gauge-based precipitation were accurately captured by BLS-based estimates. BLS method outperformed satellites on CC, RMSE, and NSE at most validation sites at low altitudes (0–1000 m). According to the independent validation, more accurate daily precipitation estimates could be obtained at more than half of the validation sites using the proposed model compared to the source datasets. The BLS-based framework considering environmental factors has the potential to improve estimates over SEC and is expected to be applied to other regions.

1. Introduction

Accurate estimation of precipitation is of critical importance for hydrometeorological applications and hazard prevention (Gao et al., 2019; Stephens & Kummerow, 2007). In general, precipitation can be derived from ground-based (e.g., rain gauges and radars) observations and satellite-based products, as well as atmospheric reanalysis products (e.g., ERA5; Tarek et al., 2020). Rain gauges and weather radars could measure precipitation with comparatively high credibility. However, rain gauges and radars are often distributed unevenly and sparsely, especially in mountainous areas (Ma et al., 2015), which leads to a deficiency of precipitation data in some regions (Gottschalck et al., 2005). Moreover, in serious natural environments, some gauges may be out of operation resulting in discontinuous rainfall measurements. Errors in the weather radar measurements can hardly be avoided, and the accuracy is influenced by weather conditions, precipitation process, and the spatial distribution of precipitation particles (Joss et al., 1998; Sokol et al., 2021). Remote sensing techniques have been applied to detect precipitation data with near-global coverage and decadal records (Ma et al., 2018; Sunilkumar et al., 2016), which could overcome the spatial and temporal restrictions to some extent. Vast satellite-based precipitation products have been released recently, such as the latest Version-7 Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA) near-real-time (3B42RT) & post-real-time (3B42V7) products (Huffman et al., 2010), precipitation products using Climate Prediction Center morphing (CMORPH) method (Joyce et al., 2004), Integrated Multi-satellite Retrievals for GPM (IMERG) product (Huffman et al., 2012), Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks—Climate Data Record (PERSIANN-CDR; Ashouri et al., 2015), and the Global Satellite Mapping of Precipitation—Moving Vector with Kalman filter (GSMaP-MVK; Ushio et al., 2009).

The performances of these widely used satellite-based and reanalysis data, however, vary with different spatial and temporal scales. The accuracy of the satellite-based products and their performance has been estimated in
previous studies (Bharti & Singh, 2015). For instance, IMERG was discovered to underestimate the light rain over the southern Tibetan Plateau, while it tends to be overestimated by 3B42V7 (Xu et al., 2017). Moreover, IMERG and 3B42V7 can yield relatively accurate precipitation data in the mid-to low-latitudes of China (Tang et al., 2016). According to Tang et al. (2020), GSMaP outperformed CMORPH in most regions of China and PERSIANN-CDR performed worse than CMORPH. However, PERSIANN-CDR could retrieve precipitation estimates with higher accuracy in winter than CMORPH. ERA5 could capture the spatial distribution and center of precipitation events over the mainland of China, but it failed to identify the precipitation extremes (Jiang et al., 2021). Moreover, reanalysis data was demonstrated to introduce substantial errors at a seasonal scale, especially in the regions characterized by complex terrains (Bhuiyan et al., 2019; Gao et al., 2017; Luo et al., 2019). In general, one satellite-based or reanalysis precipitation product does not overwhelmingly outperform the others due to different retrieving and analyzing algorithms. Therefore, to overcome the restraints of these rainfall products and promote their advantages, multi-source datasets, such as satellite-based data, ground-based data, and atmospheric reanalysis data, need to be blended (e.g., Bhuiyan et al., 2020; Ma et al., 2021). Furthermore, taking account of the environmental factors is proved to be effective to enhance the accuracy of rainfall estimates (Bharti & Singh, 2015; Tang et al., 2018; Xu et al., 2017). For instance, Bhuiyan et al. (2018) reported that the errors of assimilated rainfall data obtained by incorporating rainfall with environmental factors can be significantly reduced. Tang et al. (2018) considered environmental factors (2-m air temperature and total precipitation water) in the rainfall assimilating model based on a deep neural network to estimate rainfall and snowfall at a global scale.

In terms of methods for assimilating multi-source datasets, numerous techniques have been developed. Typically, satellite-based rainfall data could be corrected based on the in-situ gauge-based data using interpolating methods, such as optimal interpolation (OI; Xie & Xiong, 2011) and kriging interpolation (Dirks et al., 1998). In this way, prior knowledge about the relationships among the variables needs to be learned. Recently, nonparametric approaches have been proposed, whereby the explicit functional form to quantify the relationship between input-output pairs is no longer required (Hill, 2011; Bhuiyan et al., 2018). The nonparametric approaches, such as Bayesian analysis and machine learning, have been increasingly applied in estimating rainfall amounts (e.g., Bhuiyan et al., 2018; Hashim et al., 2016; Ouallouche et al., 2018; Pham et al., 2020; Ziarh et al., 2021). For instance, a two-stage blending (TSB) method based on a Bayesian theory was proposed to combine satellite- and gauge-based precipitation data, which can eliminate the potential negative effects from the individuals with poor quality (Ma et al., 2021). Specifically, nonparametric machine learning techniques, with strong abilities to model complicated relationships between massive input and output data pairs, have attracted more and more attention (Aswin et al., 2018; LeCun et al., 2015; Wu et al., 2020). Aroui et al. (2015) launched a precipitation product based on remote sensing rainfall data using the artificial neural network (ANN) technique. Wu et al. (2020) proposed a spatiotemporal deep fusion method, that is, a convolutional neural network and long-short-term memory network (CNN-LSTM), to merge satellite and gauge rainfall during 2001 and 2005 over China. However, existing machine learning methods need to overcome the time-consuming problems and enhance computational efficiency (Chen & Liu, 2017; Schmidhuber, 2015).

Broad learning system (BLS; Chen & Liu, 2017), as a nonparametric machine learning method, originating from the random vector functional-link neural network (RVFLNN; Chen & Wan, 1999), is characterized by high efficiency and is exempt from explicit functional forms. Compared to deep learning, the broad learning system does not need to train the stacks of hierarchical layers which leads to expensive computational cost (Schmidhuber, 2015), because BLS can be trained incrementally based on the information from the trained architecture (Kuok & Yuen, 2020). BLS has been successfully applied in the fields of facial expression recognition, fault diagnosis, and classification of the hyperspectral image, etc (Zhang et al., 2018; X. Zhao et al., 2019; G. Zhao et al., 2021). Nevertheless, the performance of this method in fusing multi-source precipitation data has not been tested. Furthermore, most previous studies mainly focused on assimilating multi-source satellite-based and ground-based precipitation data, but meteorological forcing factors, such as temperature, wind speed, and specific humidity, have not been paid enough attention. Simultaneously, traditional methods for merging multi-source data often only explored the temporal correlation relationship rather than considering the spatial variation of precipitation. According to characteristics of precipitation from China meteorological administration (CMA), the spatial distribution of precipitation over southeast China varies with latitudes and longitudes. Thus, geographic location in terms of latitudes and longitudes, which can provide vital information for building the input-output
relationship in the blending model, should also be considered. The performance of the assimilating model on the rainfall representation at both temporal and spatial scales should be sufficiently evaluated.

Therefore, this study aims at investigating the assimilation of multi-source precipitation data combined with environmental factors based on a new machine learning model. That is, a BLS framework is employed to integrate environmental factors (i.e., temperature, specific humidity, wind speed, and location) with satellite- and gauge-based precipitation data to yield an improved precipitation dataset at ungauged areas over SEC. The performance of the BLS-based assimilating model was evaluated in terms of a leave-one-year-out cross validation (LOYOCV) method and spatiotemporally independent validation. In both validation processes, the temporal trend of daily precipitation estimates against gauge-based daily precipitation was compared. Besides, the BLS-based average accumulative annual precipitation estimates corresponding to gauges at different altitudes were computed. The average daily precipitation estimates from six satellites and the BLS model were compared with the gauge-based data at 28 independent validation sites to test the performances of the proposed model. The capability for capturing accurate estimates of the proposed assimilating model was compared with support vector machine (SVM; Joachims, 1998).

2. Study Area and Data Sources

2.1. Study Area

As shown in Figure 1, Southeast China (SEC) extending from 15° to 35°N and from 105° to 125°E is selected as the study area, where the subtropical monsoon climate dominates. It is warm and humid in summer, and 60%–85% of the annual total precipitation of China concentrates in this area (Chen et al., 2011). From Figure 1b, the topography varies significantly in the study area, where the elevation ranges from 0 m (east and south of SEC) to 4,885 m (west of SEC). The spatial variations of rainfall and temperature in SEC share similar patterns, both of which increase from northeast to southeast of this region (Gao et al., 2008). Due to the monsoon climate and mountainous topography, disasters including floods and landslides frequently occur in this densely populated area during rainy seasons (Tang et al., 2017).
2.2. Multi-Source Precipitation Datasets

Six typical satellite-based precipitation products (3B42V7, 3B42RT, IMERG, CMORPH, GSMaP-MVK, and PERSIANN-CDR) were adopted as the source precipitation datasets. TRMM TMPA 3B42V7 and 3B42RT were employed. They are obtained based on the TMPA algorithm, through which the precipitation estimates could be retrieved from Geostationary infrared data and low earth orbiting Microwave measurements adjusted by gauge observations (Huffman et al., 2007; Kumar et al., 2021). Both two products have a resolution of 0.25° in the spatial domain and daily in the temporal domain. 3B42RT dataset covers a broader space (60°S – 60°N) compared with 3B42V7 (50°S – 50°N). The daily 3B42V7 and 3B42RT precipitation datasets were acquired from Goddard Earth Sciences Data and Information Services Center (GES DISC, https://disc.gsfc.nasa.gov/). The near-real-time Late Run Version six of IMERG precipitation product was used in this study. IMERG precipitation dataset with a fine spatial resolution (0.1° × 0.1°) and temporal interval (30 min) was generated by inter-calibrating, merging, and interpolating multiple satellite estimates combined with gauge observations (Huffman et al., 2019). IMERG precipitation product has a spatial coverage from 90°S to 90°N which is broader than that of TMPA products. IMERG precipitation datasets are also available from GES DISC. Version 1 of CMORPH satellite-gauge blended product (CMORPH-BLD, hereafter referred to as CBLD) with a spatial resolution of 0.25° × 0.25° and daily temporal scale was employed. It was derived from PMW-IR observations by using a morphing technique (Joyce et al., 2004), with global coverage, and can be accessed from 1 January 1998 to the present. The standard version seven of GSMaP-MVK with daily scale was used in this study, which was released by the Japan Aerospace Exploration Agency (JAXA). GSMaP-MVK was developed based on the PMW-IR algorithm (Kubota et al., 2020). It is available since 1 March 2014 with a spatial resolution of 0.1° × 0.1°. PERSIANN-CDR (hereafter referred to as PCDR) was obtained by assimilating the PERSIANN precipitation data and the Global Precipitation Climatology Project (GPCP) monthly product based on adaptive ANN (Ashouri et al., 2015). PCDR data at 0.25° and daily scale was used in this study.

The daily precipitation data recorded by a network of gauges was used as the reference data in this study. This dataset is examined using extreme values check, internal and spatial consistency check (Ren et al., 2010), and then released to the public by China Meteorological Data Service Center (CMDC). The gauges that had the records with null values for more than 30 days are excluded. Moreover, rainfall amounts recorded as little rain by the rain gauges are estimated to be 0 in this study. To match the temporal records of satellite- and gauge-based data, the data during 1 March 2014 and 31 December 2017 was selected. As some satellite-based products such as 3B42V7, IMERG, CBLD, and PCDR were adjusted by the monthly gauge-based GPCP product (Huffman & Bolvin, 2015), the international exchange gauge stations used in GPCP product over SEC (around 60 sites) need to be excluded. Thus, there are 279 remaining gauges in total in the study area as shown in Figure 1b. To keep the temporal consistency between gauge- and satellite-based data, the time for the gauge data was converted into UTC. Furthermore, to keep consistent with the format of gauge data (in point form), the inverse distance weighting (IDW; Bartier & Keller, 1996) method was employed to retrieve the rainfall values from these gridded satellite-based products at the corresponding locations of gauges.

2.3. Environmental Factors

The environmental factors including daily mean temperature, daily mean wind speed, near-surface specific humidity, latitude, and longitude during 1 March 2014 and 31 December 2017 were used in this study. The datasets of average daily temperature, average daily wind speed, and locations of each rain gauge can be obtained from CMDC and their spatiotemporal resolutions are consistent with those of gauge data. Near-surface specific humidity (hereafter referred to as specific humidity) dataset with 0.1° spatial resolution and 3-hourly time interval was obtained from A Big Earth Data Platform for Three Poles (He et al., 2020; Yang et al., 2010; Yang & He, 2018). Note that the original specific humidity dataset needs to be adjusted to daily data in point format using the IDW method, too.

3. Methods

3.1. The Blending Framework Based on BLS

In this study, a new blending framework based on BLS (Chen & Liu, 2017) is proposed, of which the functions are designed as follows. Let \( \Gamma(\cdot) \) represent the unknown mapping relationship between input and output variables:
where \( G \equiv [g_1, g_2, \ldots, g_{N_g}]^T \in \mathbb{R}^{N_g \times 1} \) and \( S = [s_1, s_2, \ldots, s_{N_s}] \in \mathbb{R}^{N_s \times N_s} \) denote gauge-based and satellite-based data combined with the environmental factors, respectively. \( N_g \) is the number of rainfall data points and \( N_s \) is the number of variables including satellite-based products and environmental factors.

The mapping relationship between \( G \) and \( S \) is modeled by BLS as shown in Figure 2. Considering \( S \) and \( G \) as training input and training output, the estimated training output \( \hat{G} \) can be obtained by Equation 2:

\[
\hat{G} = AW
\]

where \( A \) is a matrix originating from \( S \), and the unknown matrix, \( W \), includes the connecting weights of the broad learning system. The transformation feature matrix, \( A \), can be calculated by the following formulas:

\[
A \equiv [F^N | E^M] \equiv [F_1, F_2, \ldots, F_N | E_1, E_2, \ldots, E_M]
\]

where the \( i \)th mapped feature is computed by \( F_i = \phi(SW_i + \beta_i) \), \( i = 1, 2, \ldots, N \), and the \( j \)th enhancement node is calculated by \( E_j = \varphi(F^N W_j + \beta_j) \), \( j = 1, 2, \ldots, M \); \( N \) and \( M \) are the number of groups of mapped feature and enhancement nodes; \( \phi(\cdot) \) and \( \varphi(\cdot) \) are the functions for the feature of mapping and nonlinear activation function, respectively. Especially, in this study, \( \tanh(\cdot) \) is selected as the activation function. The form of \( \phi(\cdot) \) has no explicit restrictions (Chen et al., 2018; Chen & Liu, 2017). \( W_j, (W_{ij}) \) and \( \beta_j, (\beta_{ij}) \) are the weights and biases generated randomly. In addition, a sparse autoencoder can be used to slightly fine-tuned the weights, \( W_{ij} \), and bias, \( \beta_{ij} \), to be exempt from the limitation of randomness nature (Chen & Liu, 2017).

Given training data, \( G \), the connecting weights, \( W \), can be obtained by ridge regression approximation (Hoerl & Kennard, 1970) of pseudoinverse as expressed by Equation 4:

\[
W = A^+ G = \left( A^T A + \lambda I \right)^{-1} A^T G
\]

in which \( \lambda \) is the regulation parameter to constrain \( W \); and \( I \) is an identity matrix. The estimated output, \( \hat{G} \), can be calculated by substituting Equation 4 and Equation 3 into Equation 2. Obviously, different estimates, \( \hat{G} \), are obtained when various combinations of \( N \) feature nodes and \( M \) enhancement nodes are given. A grid search method is performed to search for \( N \) and \( M \) with one searching step (Chen et al., 2018). When \( N \) and \( M \) reach the specified number, the training process is completed with \( N \times M \) estimates. Usually, the minimum root-mean-square error (RMSE) between the estimates and the training reference data, \( G \), is employed to determine the optimal combination of \( N \) and \( M \) for this network.

Furthermore, the data fitting capability of BLS can be improved by adding new incremental feature mapping nodes and enhancement nodes without retraining the original BLS (Chen et al., 2018; Kuok & Yuen, 2020). Let \( N_\lambda \) and \( M_\lambda \) be the number of additional feature mapping groups and the number of additional enhancement nodes, respectively. Therefore, the total numbers of feature mapping groups and enhancement nodes are \( N' = N + N_\lambda \) and \( M' = M + M_\lambda \). The updated \( A \) of incremental BLS can be expressed by:

\[
A' \equiv [A | F^{(N_\lambda)} | E^{(M_\lambda)} | E^{(N_\lambda)}]
\]
where \( F^{(N_A)} = [F_{N_A+1}, F_{N_A+2}, \ldots, F_{N_A+N_A}] \) and \( E^{(M_A)} = [E_{M_A+1}, E_{M_A+2}, \ldots, E_{M_A+M_A}] \) and \( E^{(N_A)} = [E_{M_A+1}, E_{M_A+2}, \ldots, E_{M_A+N_{EF}}] \). \( E^{(N_A)} \) corresponds to the enhancement nodes due to additional \( N_A \) feature mapping groups and \( N_{EF} \) is the number of the enhancement nodes induced by additional mapped feature groups. All the additional mapping feature groups and enhancement nodes can be computed by \( F_i \) and \( E_j \), respectively.

The pseudoinverse of \( A' \) is expressed as:

\[
(A')^+ = \begin{bmatrix} A^* - D B \\ B \end{bmatrix}
\]

(6)

where

\[
D = A^* \left[ F^{(N_A)} | E^{(M_A)} | E^{(N_A)} \right]
\]

(7)

\[
B = \begin{cases} 
C^+, & C \neq 0 \\
(I + D^T D)^{-1} D^T A^*, & C = 0 
\end{cases}
\]

(8)

\[
C = F^{(N_A)} | E^{(M_A)} | E^{(N_A)} - AD
\]

(9)

The updating weights of the incremental BLS can be obtained by

\[
W' = (A')^+ G
\]

(10)

The training output of the incremental BLS can be computed by \( G = A' W' \). For the BLS with additional mapping feature groups and/or enhancement nodes, RMSE between training output and training gauge data (referred as to training error) is still computed. When training error meets RMSE is less than training error tolerance (a prescribed number, such as 10\(^{-3}\)) or the threshold numbers of \( N \) and \( M \) are reached, the incremental process will be terminated.

### 3.2. Methods for Evaluating the BLS-Based Framework

To evaluate the performance of the proposed BLS framework for assimilating multi-source datasets, LOYOCV method (Schepen & Wang, 2014) was adopted. The data corresponding to the validation sites from 1 year was used to validate and the data from the remaining years were employed to train the proposed BLS model, and this procedure was repeated several times. Specifically, the datasets including six satellite-based products, gauge-based data as well as five environmental factors in the period of 1 March 2014 and 31 December 2017 were split into two parts (training and validation data, as shown in Figure 3). Training data corresponding to all the stations over SEC from 3 years was used to train the BLS-based model, and validation data corresponding to the sites in the remaining 1 year was employed to verify the trained BLS. Note that the total number of gauges over SEC in each year is 279 as shown in Figure 1b. The training and validation data are independent of each other at a temporal scale. The experiment based on LOYOCV was repeated four times during the study period.

To verify the transitivity of the proposed assimilating model, another validating experiment was designed. Two gauge stations were randomly selected in each province of SEC so that there were 28 stations (referred to as validation sites) selected in this experiment. The specific information about these validation sites is presented in Table 1. The data corresponding to the 251 remaining stations (training sites) from 1 March 2014 to 31 December 2016 over SEC was referred to as training data which was used to train the BLS model. The data corresponding to the 28 validation sites in 2017 was referred to as validation data to verify the trained model. The training and validation data in this experiment are not only temporally but also spatially independent with each other. Therefore, the 28 validation sites could be considered as ungauged areas to validate the performance of the BLS model.

Four statistical metrics, that is, Pearson's correlation coefficient (CC), RMSE, mean absolute error (MAE), and Nash-Sutcliffe coefficient of efficiency (NSE), were employed to assess the precipitation estimates against gauge.
data in two experiments presented in Table 2. In these expressions, $g_p$ and $g_s$ are the $p$th sample of estimated rainfall, $\hat{G}$, and gauge-based validation rainfall data, $\bar{G}$, respectively. $\hat{\bar{G}}$ and $\bar{\bar{G}}$ are the mean of $\hat{G}$ and $\bar{G}$.

4. Performance of BLS-Based Assimilating Model

4.1. Evaluation Using Estimates in LOYOCV

The proposed BLS model was evaluated using LOYOCV during the period from 1 March 2014 to 31 December 2017. Six satellite-based datasets (3B42V7, 3B42RT, IMERG, CBLD, GSMaP, and PCDR) combined with five

| Name  | Location          | Province | Name  | Location          | Province |
|-------|-------------------|----------|-------|-------------------|----------|
| Site 1| (107.32°E, 34.03°N) | Shanxi   | Site 15| (116.78°E, 32.43°N) | Anhui    |
| Site 2| (111.65°E, 33.78°N) | Henan    | Site 16| (120.57°E, 32.37°N) | Jiangsu  |
| Site 3| (109.15°E, 33.43°N) | Shanxi   | Site 17| (117.85°E, 30.98°N) | Anhui    |
| Site 4| (114.52°E, 33.78°N) | Henan    | Site 18| (119.70°E, 30.22°N) | Zhejiang |
| Site 5| (105.98°E, 31.58°N) | Chongqing| Site 19| (119.93°E, 28.47°N) | Zhejiang |
| Site 6| (111.83°E, 31.80°N) | Hubei    | Site 20| (118.32°E, 27.05°N) | Fujian   |
| Site 7| (113.95°E, 30.90°N) | Hubei    | Site 21| (116.02°E, 26.48°N) | Jiangxi  |
| Site 8| (107.73°E, 29.85°N) | Chongqing| Site 22| (117.42°E, 25.30°N) | Fujian   |
| Site 9| (111.22°E, 28.38°N) | Hunan    | Site 23| (113.42°E, 24.18°N) | Guangdong|
| Site 10| (114.78°E, 28.40°N) | Jiangxi  | Site 24| (107.57°E, 23.28°N) | Guangxi  |
| Site 11| (108.25°E, 27.95°N) | Guizhou  | Site 25| (110.08°E, 23.40°N) | Guangxi  |
| Site 12| (112.40°E, 26.42°N) | Hunan    | Site 26| (112.78°E, 22.25°N) | Guangdong|
| Site 13| (106.63°E, 26.13°N) | Guizhou  | Site 27| (109.83°E, 19.03°N) | Hainan   |
| Site 14| (118.78°E, 34.08°N) | Jiangsu  | Site 28| (110.47°E, 19.23°N) | Hainan   |
Table 2

| Statistical metrics                        | Expression                                                                 | Ranges          |
|--------------------------------------------|---------------------------------------------------------------------------|-----------------|
| Correlation coefficient (CC)               | \( CC = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \sum_{j=1}^{N} (y_j - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{j=1}^{N} (y_j - \bar{y})^2}} \) | \([-1, 1]\)     |
| Root-mean-square error (RMSE)              | \( RMSE = \sqrt{\frac{\sum_{i=1}^{N} (x_i - y_i)^2}{N}} \)                | \([0, +\infty)\) |
| Mean absolute error (MAE)                  | \( MAE = \frac{\sum_{i=1}^{N} |x_i - y_i|}{N} \)                          | \([0, +\infty)\) |
| Nash-Sutcliffe coefficient of efficiency (NSE) | \( NSE = 1 - \frac{\sum_{i=1}^{N} (x_i - y_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \) | \((-\infty, 1]\) |

Figure 4. The broad learning system-based daily precipitation against gauge-based daily precipitation corresponding to the validation site where root square mean error of estimates at (a) the maximal level, (b) the average level, and (c) the minimal level, and the corresponding scatter plots in (d), (e) and (f).

Environmental factors (i.e., temperature, specific humidity, wind speed, latitude, and longitude) were employed in the LOYOCY process. The comparison between daily precipitation estimates and gauge-based observations was evaluated. BLS-based average accumulative annual precipitation corresponding to gauges at different altitudes was assessed.

4.1.1. BLS-Based Precipitation Estimates at Temporal Scale

The 279 sites in 2017 over SEC as presented in Figure 1b were referred to as validation sites in the LOYOCY process. Figure 4 demonstrates the time series and scatter plots of BLS-based daily precipitation against gauge-based observations at the validation sites with the maximal RMSE, average RMSE, and minimal RMSE of estimates. The daily estimates with maximal RMSE (18.56 mm) corresponded to validation sites at the south of Guangxi province where rainstorm events (daily precipitation amount more than 50 mm) often occurred. According to the gauge-based data, 18 rainstorm events were recorded at this site. BLS could estimate precipitation more accurately when daily precipitation was less than 50 mm compared with those during rainstorm events. Especially, the temporal trend of gauge-based daily precipitation at this site could be simulated by BLS-based estimates. Variations of rainfall amounts from BLS in warm and cold seasons were notable, which was consistent with those from gauges. Moreover, the scatter plot in Figure 4d also demonstrates the capability of BLS for capturing daily precipitation less than 50 mm. BLS-based estimates linearly agreed with ground-based observations with a CC value of 0.754. In terms of the other three statistical metrics, IMERG performed the best compared to other methods. As for estimates with the average (5.95 mm) and minimal (1.53 mm) RMSEs, they originated from the validation sites located at the east of Guangxi and the northwest of Henan provinces, respectively. BLS still underestimated daily rainfall amount during rainstorm events as shown in Figure 4b, while it can accurately estimate rainfall amounts during the non-rainstorm events as presented in Figure 4c. Estimates at both sites can match with the temporal trends from gauges well, especially for those in Henan province. The corresponding scatter plots are shown in Figures 4e and 4f, in which serval outliers could be observed especially in Figure 4e. Except for that, more points at sites in Henan province are scattered on both sides of the diagonal, indicating that BLS-based estimates at this site agreed with those from gauges better in contrast to those in Guangxi province. Actually, CC values at the
three sites are 0.754 (Guangxi), 0.769 (Guangxi), and 0.962 (Henan). That is, more accurate estimates can be obtained when fewer rainstorm events are recorded at gauges. BLS is prone to yield accurate daily precipitation when daily precipitation is less than 50 mm in SEC.

Figure 5 shows scatter plots of BLS-based estimates of daily precipitation against gauge-based daily precipitation at 28 out of 279 validation sites (as presented in Table 1) in 2017 over SEC. Similar outliers with those shown in Figures 4d–4f were observed in this figure, where the BLS model yielded obvious overestimated/underestimated daily precipitation ranging between 0.1 and 1 mm as well as 100 and 200 mm. Except for these points, the other points were scattered along the diagonal, indicating that BLS-based estimates at the 28 validation sites could linearly match with gauge-based daily precipitation whose values range between 1 and 100 mm. The mean CC value of BLS-based estimates at 28 sites was 0.857, followed by CBLD (0.848), IMERG (0.736), 3B42V7 (0.687), and 3B42RT (0.672), GSMaP (0.635), and PCDR (0.560). In terms of RMSE, the BLS scheme can obtain the daily precipitation with the smallest errors compared with the satellite products. Moreover, the largest mean NSE at 28 sites with a value of 0.717 was originated from the BLS method while the six satellites exhibited relatively smaller mean NSEs which were 0.699 (CBLD), 0.345 (3B42V7), 0.293 (IMERG), 0.220 (PCDR), 0.203 (3B42RT), and −0.152 (GSMap), respectively. Overall, in contrast to source data, the proposed model yielded the results with the relatively better three statistical metrics, which proves that it could obtain more accurate daily precipitation over SEC.

4.1.2. BLS-Based Precipitation Estimates at Different Altitudes

Altitudes of 279 validation sites were classified into four divisions according to the values at the 25th percentile, 50th percentile, and 75th percentile of altitudes. That is, four divisions, 0–57.68 m, 57.68 m–142 m, and 142 m–381.33 m, as well as higher than 381.33 m, were included. Figure 6 demonstrates the accumulative annual precipitation corresponding to each site located at different altitudes including four categories of altitudes in 2017 over SEC. The thick line indicates average accumulative annual precipitation from BLS-based estimates. The maximal accumulative annual precipitation (3,480.2 mm) in Figure 6a was recorded by the gauge located at the south of Guangxi province with an altitude of 56.8 m. Another gauge located at the border of Jiangxi and Hubei provinces (altitude is 99.9 m) measured the maximal accumulative annual precipitation with 2,251.9 mm as shown in Figure 6b. Where the altitudes ranged between 142 and 381.33 m, the accumulative annual precipitation recorded at most of the gauges ranged between 1,000 and 2,000 mm. Nevertheless, where altitudes were above 381 m, smaller accumulative annual precipitation amounts at most of the gauges were observed, as shown in Figure 6d. Almost all accumulative precipitation in the four subfigures increases sharply from the 100th day to the 280th day, which exactly represented the characteristics of precipitation in the study area. This temporal trend was depicted by the average BLS-based accumulative precipitation as marked by the thick lines in these subfigures. Moreover, the estimated average accumulative annual precipitation in Figure 6d is the smallest (1,302.5 mm), compared with the other three values in Figures 6a–6c, which is consistent with the gauge observations. All the four thick curves are at the average level of all gauge-based accumulative curves at different altitudes in four divisions, indicating that estimates from BLS are reasonable.

The altitudes in SEC are in general low (0–1,000 m), and high altitudes (higher than 1,000 m) are mainly located at the west of the study area. According to previous studies, precipitation amounts may change with the varying altitudes (Ma et al., 2018). Therefore, it is essential to evaluate the capability of the proposed BLS framework for capturing accurate precipitation at different altitudes. Figure 7 compares the precipitation estimates from six satellite-based methods and the BLS-based model (with environmental factors) with the gauge-based data at low altitudes over SEC in 2017. Generally, significant variations of distribution of these four metrics from the six products were observed. Especially, the smaller CCs from all the products were found in the north and
Moreover, two TRMM satellites, GSMaP, and PCDR yielded CC with smaller values at most of the sites along the seashores of SEC, while PCDR performed the worst over these regions. IMERG, CBLD, and BLS schemes obtained relatively larger CCs over these locations. The spatial distributions of CCs from BLS and CBLD were similar with each other, but BLS obtained the largest CC values at 54.6% of gauge stations (31.7% gauges for CBLD), indicating the BLS-based method performed better on CC at most gauge stations in contrast to six original satellites. In terms of RMSE, most larger errors from six satellite schemes were distributed over the south and middle of SEC, while lower values were observed from the proposed model in this area. Larger MAEs from the seven methods, especially GSMaP and PCDR, were found in Guangdong, Guangxi, Fujian, and Jiangxi provinces followed by two TRMM satellites and IMERG. That is, CBLD and BLS were capable of diminishing more errors in these areas compared with the other five approaches. As for NSE, the BLS approach obviously performed the best among the seven methods due to wide distributions of NSE with larger values. NSEs from BLS at 150 sites were larger than those from CBLD (83 sites), IMERG (9 sites), 3B42V7 (4 sites), 3B42RT (1 site), and GSMaP (0 site) and PCDR (0 site), which means the assimilated precipitation from the proposed scheme can simulate the ground observations well at around 57.3% of sites. Overall, GSMaP and PCDR are not recommended to be applied to the SEC area with low altitudes compared with other methods.

Figure 6. Accumulative precipitation recorded by gauges located at different altitudes including (a) <25th percentile of altitudes, (b) between 25th and 50th percentile of altitudes, (c) between 50th and 75th percentile of altitudes, and (d) above 75th percentile of altitudes in 2017 over southeast China. Note that the thick line indicates average accumulative precipitation from broad learning system over validation sites at different altitudes.
The spatial distribution of daily statistical metrics at high altitudes was also compared (the corresponding figures are not shown here). In general, in contrast to six satellites, the BLS-based scheme presented similar performance with CBLD on CC, RMSE, and NSE at high altitudes over SEC.

4.2. Evaluation Using Estimates in Independent Validation

Precipitation estimations were obtained by using validation data from six satellite-based products incorporating five environmental factors corresponding to 28 validation sites (shown in Table 1) in 2017 as the input of the trained BLS model during the temporally and spatially independent validation. Figure 8 shows BLS-based daily estimates against gauge-based daily precipitation at 28 selected validation sites in 2017 over SEC, and the corresponding scatter plots are shown accordingly. In general, estimates corresponding to daily precipitation less than 50 mm from BLS were more consistent with those from gauge-based data. BLS yielded estimated precipitation with better performance at sites in provinces along seashores, such as Sites 18, 22, 25, and 28, compared with those located at inland areas as shown in Figure 8. A similar performance was also observed from the corresponding scatter plots, in which the estimated precipitation from BLS at these four sites agreed with the ground-based observations with CC values of 0.875, 0.962, 0.943, and 0.930, respectively. NSE values of estimates from BLS at the four sites were 0.723, 0.918, 0.885, and 0.860, respectively. Rainstorm events with daily precipitation more than 50 mm often occurred in the warm season, when relatively slight underestimations were obtained using the BLS method. However, when the gauge-based daily precipitation was less than 50 mm, the BLS scheme can capture relatively accurate estimates, for example, at Sites 4, 11, 21, and 22. Some outliers were observed such as at Site 9 and Site 15 as shown in the scatter plots. Most of these outliers corresponded to the daily precipitation from gauges with values more than 50 mm. Simultaneously, significant overestimations from BLS at Site 14 corresponding to gauge-based precipitation were found. Except for that, the estimates from the BLS scheme were scattered on both sides of the diagonal, indicating that the BLS model can yield relatively accurate estimates when the corresponding daily gauge-based precipitation is less than 50 mm.

Figure 9 shows the spatial distribution of daily statistical metrics of estimates from six satellite- and BLS-based methods at 28 validation sites over SEC in 2017. Except for CCs from GSMaP and PCDR, those from the other
Figure 8. Time series and scatter plots of broad learning system-based estimated daily precipitation against gauge-based daily precipitation at each validation site in 2017 over southeast China.

Figure 9. Spatial distribution of daily statistical metrics of precipitation estimates from six satellites and broad learning system against gauge-based data corresponding to validation sites using independent validation over SEC in 2017.
methods were more than 0.4, but more CCs with relatively larger values were observed from BLS-based estimates in contrast to the source datasets. Moreover, the minimum CC from BLS was 0.72, which is larger than those of other methods. As for RMSE, relatively larger RMSEs were observed in Hainan province, where BLS- and BCLD-based methods obtained the estimates with the smaller RMSE at one of these two sites compared to the other five approaches. Moreover, the BLS-based method yielded the precipitation with the smallest RMSE at 15 validation sites, but the satellite-based methods, especially 3B42RT and GSMaP, obtained the estimates with more errors at these stations. In terms of MAE, the spatial distributions from these seven methods were similar to those of RMSE. In terms of NSE, BLS performed the best at 15 sites, followed by CBLD at the other 13 sites. GSMaP did not simulate the gauge-based precipitation well over northwest and northeast of SEC, but IMERG and 3B42RT cannot obtain larger NSE at the sites over northeast and southeast of SEC.

Figure 10 shows the daily average-areal precipitation from six satellite- and BLS-based approaches against gauge-based observations at validation sites over SEC in 2017. All the methods overall could capture the temporal trends of gauge-based daily precipitation. Particularly, compared with the other five schemes, 3B42RT and PCDR showed the worst performance in January and February, as they underestimated the daily average-areal precipitation, further, to fail to capture the temporal trend of gauge observations in these 2 months. In March, obvious overestimations were observed from 3B42V7, IMERG, and GSMaP, especially when the daily average-areal gauge-based precipitation was less than 10 mm. Moreover, 3B42RT, IMERG, and PCDR also overestimated the precipitation in June. BLS and CBLD obtained relatively accurate estimates during this period. All the approaches underestimated the precipitation when the daily average-areal gauge-based precipitation was the maximal this year. BLS and CBLD could yield more accurate estimates in the remaining days of this year in contrast to other original satellites. The corresponding scatter plots are shown in Figures 10h–10n. In general, average estimates from BLS agreed with ground observations the best with the maximal CC of 0.951, followed by CBLD (0.949), IMERG (0.865), 3B42V7 (0.862), 3B42RT (0.840), GSMaP (0.816), and PCDR (0.793). To be specific, when the daily average-areal gauge-based precipitation amount was, the more accurate BLS-based estimates were. The proposed model has the potential to yield improved daily average-areal precipitation over SEC.

According to the performance of the BLS model in independent validation, the proposed model could obtain the average daily precipitation with relatively fewer errors over SEC compared with the source datasets.

5. Discussion

5.1. Effectiveness of Environmental Factors in BLS-Based Model

The effectiveness of environmental factors in the BLS-based model was assessed by four average daily statistical metrics in Table 3, where estimates with factors and no factors indicate the assimilating model incorporating
Table 3
Average Daily Statistical Metrics of Precipitation Estimates Corresponding to 279 Validation Sites in Each Year Using Leave-One-Year-Out Cross Validation

| Test period | Products    | CC     | RMSE (mm) | MAE (mm) | NSE  |
|-------------|-------------|--------|-----------|----------|------|
| 2014        | 3B42V7      | 0.711  | 8.997     | 3.584    | 0.391|
|             | 3B42RT      | 0.683  | 9.598     | 3.859    | 0.299|
|             | IMERG       | 0.733  | 8.997     | 3.428    | 0.386|
|             | CBLD        | 0.823  | 6.658     | 2.577    | 0.655|
|             | GSMaP       | 0.685  | 10.049    | 3.607    | 0.138|
|             | PCDR        | 0.551  | 10.333    | 4.608    | 0.210|
|             | Estimates (no factors) | 0.825  | 6.74      | 2.636    | 0.657|
|             | Estimates (with factors) | 0.833  | 6.580     | 2.595    | 0.673|
| 2015        | 3B42V7      | 0.690  | 9.075     | 3.639    | 0.316|
|             | 3B42RT      | 0.655  | 10.220    | 4.008    | 0.109|
|             | IMERG       | 0.707  | 9.181     | 3.504    | 0.266|
|             | CBLD        | 0.817  | 6.455     | 2.432    | 0.637|
|             | GSMaP       | 0.651  | 10.297    | 3.627    | −0.115|
|             | PCDR        | 0.543  | 10.126    | 4.560    | 0.170|
|             | Estimates (no factors) | 0.824  | 6.392     | 2.428    | 0.655|
|             | Estimates (with factors) | 0.828  | 6.318     | 2.488    | 0.666|
| 2016        | 3B42V7      | 0.680  | 9.816     | 3.902    | 0.304|
|             | 3B42RT      | 0.647  | 11.079    | 4.301    | 0.067|
|             | IMERG       | 0.714  | 9.863     | 3.628    | 0.273|
|             | CBLD        | 0.808  | 6.983     | 2.605    | 0.607|
|             | GSMaP       | 0.659  | 11.164    | 3.971    | −0.370|
|             | PCDR        | 0.555  | 10.488    | 4.670    | 0.224|
|             | Estimates (no factors) | 0.811  | 7.098     | 2.604    | 0.611|
|             | Estimates (with factors) | 0.819  | 6.822     | 2.645    | 0.639|
| 2017        | 3B42V7      | 0.672  | 8.639     | 3.479    | 0.325|
|             | 3B42RT      | 0.658  | 9.308     | 3.688    | 0.195|
|             | IMERG       | 0.723  | 8.362     | 3.256    | 0.347|
|             | CBLD        | 0.819  | 6.087     | 2.314    | 0.645|
|             | GSMaP       | 0.622  | 10.557    | 3.576    | −0.465|
|             | PCDR        | 0.549  | 9.416     | 4.168    | 0.203|
|             | SVM (with factors) | 0.825  | 6.094     | 2.269    | 0.660|
|             | Estimates (no factors) | 0.827  | 6.028     | 2.334    | 0.659|
|             | Estimates (with factors) | 0.831  | 5.953     | 2.375    | 0.672|

Note. The number in bold indicates the best statistical indicator.

Overall, the BLS model combined with no factors performed better on almost all statistical indicators compared with the six satellite-based techniques, however, it obtained more inaccurate estimates than the BLS model incorporating five factors did in each year. Especially, CC of estimates with no factors were 0.825, 0.824, 0.811, 0.827 in 2014, 2015, 2016, and 2017, respectively, which were all slightly smaller than those of estimates with factors in the corresponding test year. Although the BLS model with no factors outperformed the scheme with factors on MAE in 2015 and 2016, better performance on RMSE and NSE from the BLS model combined with factors could be observed during each test period. The better performance of the BLS model with factors demonstrated that five environmental factors added into the BLS model could enhance the ability of the model for capturing more accurate precipitation.

Furthermore, the effectiveness of location information in the trained model was also tested. Based on the latitudes and longitudes over SEC, a latitude and longitude network with a spatial resolution of 0.1° was obtained. The daily average-areal precipitation and daily average-areal meteorological factors in warm and cold seasons, as well as in 2016 combined with the network of location were set as the input of the trained BLS model (this model has been obtained in the LOYOCV process). That is, the input matrix was defined as \( \mathbf{S} = \{V_7, RT, IM, CMO, GP, PR, SH, T, W, Lat, Lon\} \), where the first nine variables are daily average-areal rainfall from 3B42V7, 3B42RT, IMERG, CBLD, GSMaP, PCDR, and daily average-areal meteorological factors corresponding to specific humidity, temperature, and wind speed, respectively. The last two variables are varying latitudes and longitudes (\( 1 \leq i, j \leq 401401 \), which is the total number of gridded points over SEC). The first nine values were specified numbers and did not change with locations. The precipitation data and meteorological data in 2016 were used as frequent heavy rain occurred in this year which resulted in more obvious characteristics of spatial pattern of rainfall in contrast to other years. The spatial distribution of estimated average daily precipitation and gauge-based average daily precipitation is shown in Figure 11. Note that if there are inaccurate measurements recorded at gauges in 2016, the gauge stations will not be shown in this figure. Estimates in the warm season from April to September (Zhong, 2020) as shown in Figure 11a were overall larger than those in the cold season (October to March; Figure 11b), which is consistent with the observations from gauges shown in Figures 11d and 11e. Moreover, precipitation amounts over Guangdong, Guangxi, Hainan, Fujian, Hunan, and Jiangxi were relatively larger with five factors and none factors, respectively. Compared with the schemes considering only the satellite products, the BLS scheme with environmental factors performed the best on the three statistical indicators in general.

The RMSE of daily BLS-based precipitation were 6.580 mm, 6.318 mm, and 5.953 mm in 2014, 2015, 2016, and 2017, respectively, which was significantly less than those from the original individuals, such as GSMaP and PCDR in each test year. Moreover, the mean value of NSE was about 0.663 using the proposed model in the LOYOCV process, which was notably larger than those from the six satellites, especially GSMaP (the smallest NSE with a value of −0.465 was observed in 2017). CC values from these seven methods in the 4 years varied from 0.543 from PCDR in 2015 to 0.833 from BLS in 2014, while the BLS method can yield the largest CC in each test year, followed by CBLD, IMERG, 3B42V7, GSMaP (3B42RT), and PCDR. RMSE and MAE fluctuated relatively drastically in these years, especially in 2015 and 2016. According to previous studies (Tao et al., 2016; X. Zhao & Niu, 2019), relatively heavy rain occurred in these 2 years compared with the other years.
than those in other provinces in SEC at the annual scale, which also can be observed in Figure 11f. Although underestimates/overestimates were presented in the northwest of SEC, south of Jiangsu and Anhui provinces, the designed scheme overall can capture the spatial characteristics of precipitation that varies with locations over SEC. These comparisons indicate the proposed model could simulate the complicated relationship between the reference precipitation data and locations. The geographical factor in terms of latitude and longitude could provide useful information for the model, and further portray the spatial pattern of average daily precipitation over SEC.

5.2. Comparison With SVM Model

An assimilating model based on SVM was employed to compare with the BLS model. The performance of SVM on four statistical indicators in 2017 was presented in Table 3. Six satellite-based products incorporated with five environmental factors were also included in the SVM model. SVM can more accurately estimate
daily precipitation with CC of 0.825, MAE of 2.269 mm, and NSE of 0.660 compared to source products. However, BLS outperformed SVM on CC, RMSE, and NSE, which means that the proposed model is capable of obtaining relatively accurate daily precipitation over SEC in contrast to SVM.

5.3. Efficiency of BLS-Based Model

If the fitting capability of the trained model is required to improve, there is no need to retrain the whole model from the beginning as it can inherit the information from the trained structure. That is, using Equations 5–10 the trained BLS model can be updated by adding incremental enhancement nodes or feature nodes without relearning, which is efficient and time-saving. The training data from 2014, 2015, 2016 and validation data in 2017 were taken as an example to demonstrate the efficiency of the incremental BLS model. 30 enhancement nodes at each were added into the trained BLS model with 9 feature mapping groups, 30 enhancement nodes and training error with the value of 0.0155. The accuracy and the computation time are listed in Table 4. The training time used to compute the additional 30 enhancement nodes was around 0.3168 s which was significantly shorter than that (0.5568 s) for computation with 9 feature mapping groups and 60 enhancement nodes. The training errors of these two scenarios are equal, which means that incremental learning can obtain accurate results. The computation was performed on the MATLAB platform using a 16.0-GB RAM computer with a 3.0-GHz Intel i7-9700 CPU processor. This characteristic of BLS is quite attractive particularly when the trained system has complicated networks and improved fitting capability requirements.

Furthermore, the efficiency of an assimilating model based on a deep neural network (DNN; LeCun et al. [2015]) was computed, which was compared with the proposed method. Two DNN-based frameworks were considered. One of them was set as architecture with fewer layers (N_{layer} = 3) denoted as DNN1, and the other was referred to as DNN2 with more layers (N_{layer} = 7). In each layer, there are 10 neurons for DNN1 and DNN2. Levenberg–Marquardt algorithm (Moré, 1978) was used to train these two DNN-based networks. For BLS, the feature mapping groups and enhancement nodes are 9 and 60, respectively. Table 5 summarizes the performance of two DNN- and BLS-based networks on training time and validation error (RMSE). The data from 2014 to 2016 was used to train these networks, and data in 2017 was adopted to validate them. The performance of DNN1 on RMSE is worse than that of DNN2 as the simple architecture of DNN1 fails to depict the complex input-output relationship. DNN2 performs similarly on RMSE with BLS, but it consumes much longer training time, indicating that BLS is more efficient than DNN.

6. Conclusions

This study investigated the performance of a BLS framework on incorporating multi-source precipitation data with environmental factors over SEC. The daily satellite-based precipitation data (3B42V7, 3B42RT, IMERG, CBLD, GSMaP, and PCDR) and daily gauge-based data combined with daily environmental factors (temperature, specific humidity, wind speed, and location) during 2014 and 2017 were included in this work. This proposed framework was evaluated using LOYOCV and independent validation and assessed in terms of four statistical metrics from different perspectives at various spatiotemporal scales. The conclusions can be summarized as follows.

1. In each year of the study period, the BLS-based precipitation estimates from the LOYOCV experiment were significantly improved with better statistical metrics (CC, RMSE, and NSE), especially larger NSE, in contrast to the source data across SEC. Estimated daily precipitation at 28 out of 279 validation sites agreed with gauge-based daily observations ranging between 1 and 50 mm better in contrast to those in other ranges. Seasonal variations of daily gauge-based precipitation were accurately captured by BLS-based estimates. The estimated amounts of average accumulative annual precipitation at validation sites at altitudes above 381 m were smaller than those at altitudes lower than 381 m. BLS method yielded the estimates with larger CC, smaller RMSE, as well as larger NSE at most sites located at low altitudes.

| N | M   | Training time (s) | Testing time (s) | Training error |
|---|-----|------------------|-----------------|---------------|
| 9 | 30  | 0.4067           | 0.0743          | 0.0155        |
| 9 | 30- > 60 | 0.3168           | 0.0390          | 0.0154        |
| 9 | 60  | 0.5568           | 0.0965          | 0.0154        |

| Model | Training time (s) | RMSE (mm) |
|-------|------------------|-----------|
| DNN1  | 35.12            | 5.969     |
| DNN2  | 58.20            | 5.957     |
| BLS   | 0.56             | 5.953     |

Table 4: Accuracy and Efficiency of BLS Model With Incremental Learning

Table 5: Performance of Assimilating Framework Based on DNN and BLS With Different Configurations
2. In the independent validation process, the temporal trend of gauge-based daily precipitation at each validation site can be simulated by BLS. Moreover, in contrast to gauge data in other ranges, BLS can yield more accurate estimates when the corresponding gauge-based daily precipitation is less than 50 mm. BLS model outperformed the six satellite products on CC, RMSE, and NSE at more than half of the validation sites.

3. BLS model considering environmental factors outperformed the scheme without environmental factors on CC, RMSE, and NSE. Spatial patterns of estimates from the only varying location as input of the trained BLS model were consistent with the spatial characteristics of gauge measurements over SEC. BLS method can simulate the complicated mapping relationship between reference gauge data and six satellite-based data combined with five environmental factors, especially the relationship between precipitation and location over SEC. The proposed model has the potential to obtain relatively accurate daily precipitation at ungauged areas over SEC.

Data Availability Statement

The data in this study are subject to third party restrictions. The data that support the findings of this study are available from the National Climate Centre in Beijing, China. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://data.cma.cn/en, accessed on 11 January 2020, with the permission of the National Climate Centre in Beijing, China.

Acknowledgments

This paper is financially supported by the Science and Technology Development Fund, Macau SAR (File no.: SKL-IOTSC-2021-2023, 00302020/A1, 0021/2020/ASC), UM Research Grant (File no.: SRI2019-00193-IOTSC, MYRG2020-00072-IOTSC), Guangdong–Hong Kong–Macau Joint Laboratory Program (Project No.: 2020B1212030009), and CORE (EF017/IOTSC-GL/2020/HKUST).

References

Ashouri, H., Hsu, K. L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, I. D., et al. (2015). PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. Bulletin of the American Meteorological Society, 96(1), 69–83.

Aswin, S., Geetha, P., & Vinayakumar, R. (2018). Deep learning models for the prediction of rainfall. In 2018 International Conference on Communication and Signal processing (ICCCSP) (pp. 0657–0661). IEEE.

Barrier, P. M., & Keller, C. P. (1996). Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). Computers & Geosciences, 22(7), 795–799. https://doi.org/10.1016/0098-3004(96)00021-0

Bharti, V., & Singh, C. (2015). Evaluation of error in TRMM 3B42V7 precipitation estimates over the Himalayan region. Journal of Geophysical Research: Atmospheres, 120(24), 12458–12473. https://doi.org/10.1002/2015jd023779

Bhuiyan, M. A. E., Nikolopoulos, E. I., & Anagnostou, E. N. (2019). Machine learning–based blending of satellite and reanalysis precipitation datasets: A multiregional tropical complex terrain evaluation. Journal of Hydrometeorology, 20(11), 2147–2161. https://doi.org/10.1175/jhm-d-19-0073.1

Bhuiyan, M. A. E., Nikolopoulos, E. I., Anagnostou, E. N., Quintana-Seguí, P., & Barella-Ortiz, A. (2018). A nonparametric statistical technique for combining global precipitation datasets: Development and hydrological evaluation over the Iberian Peninsula. Hydrology and Earth System Sciences, 22(2), 1371–1389.

Bhuiyan, M. A. E., Yang, F., Biswas, N. K., Rahat, S. H., & Neelam, T. J. (2020). Machine learning-based error modeling to improve GPM IMERG precipitation product over the Brahmaputra river basin. Forecasting, 26, 245–266. https://doi.org/10.3390/forecast2020014

Chen, C. P., & Liu, Z. (2017). Broad learning system: An efficient and effective incremental learning system without the need for deep architecture. IEEE Transactions on Neural Networks and Learning Systems, 29(1), 10–24.

Chen, C. P., Liu, Z., & Feng, S. (2018). Universal approximation capability of broad learning system and its structural variations. IEEE Transactions on Neural Networks and Learning Systems, 30(4), 1191–1204.

Chen, C. P., & Wan, J. Z. (1999). A rapid learning and dynamic stepwise updating algorithm for flat neural networks and the application to time-series prediction. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(1), 62–72. https://doi.org/10.1109/3477.740166

Chen, W., Jiang, Z., Li, L., & Yiou, P. (2011). Simulation of regional climate change under the IPCC A2 scenario in southeast China. Journal of Hydrometeorology, 12(2), 491–507. https://doi.org/10.1175/2010jhm1395.1

Dirks, K. N., Hay, J. E., Stow, C. D., & Harris, D. (1998). High-resolution studies of rainfall on Norfolk Island: Part II: Interpolation of rainfall data. Journal of Hydrology, 208(3–4), 187–193. https://doi.org/10.1016/S0022-1694(98)00155-3

Gao, L., Bernhardt, M., Schulz, K., & Chen, X. (2017). Elevation correction of ERA-Interim temperature data in the Tibetan Plateau. International Journal of Climatology, 37(9), 3540–3552. https://doi.org/10.1002/joc.4935

Gao, L., Zhang, L., Li, X., & Zhou, S. (2019). Evaluating metropolitan flood coping capabilities under heavy storms. Journal of Hydrologic Engineering, 24(6), 05019011. https://doi.org/10.1061/(asce)he.1943-5584.0001793

Gao, X., Shi, Y., Song, R., Giorgi, F., Wang, Y., & Zhang, D. (2008). Reduction of future monsoon precipitation over China: Comparison between a high resolution RCM simulation and the driving GCM. Meteorology and Atmospheric Physics, 100(1), 73–86. https://doi.org/10.1007/s00703-008-0296-5

Gotovschalck, J., Meng, J., Rodell, M., & Houser, P. (2005). Analysis of multiple precipitation products and preliminary assessment of their impact on global land data assimilation system land surface states. Journal of Hydrometeorology, 6(5), 573–598. https://doi.org/10.1175/jhm437.1

Hashim, R., Roy, C., Motamedi, S., Shamsabirband, S., Pirković, D., Gacic, M., & Lee, S. C. (2016). Selection of meteorological parameters affecting rainfall estimation using neuro-fuzzy computing methodology. Atmospheric Research, 171, 21–30. https://doi.org/10.1016/j.atmosres.2015.12.002

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., & Li, X. (2020). The first high-resolution meteorological forcing dataset for land process studies over China. Scientific Data, 7(1), 1–11. https://doi.org/10.1038/s41597-020-0369-y

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. Journal of Computational & Graphical Statistics, 20(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

https://doi.org/10.1029/2021EA002043
Ushio, T., Sasashige, K., Kubota, T., Shige, S., Okamoto, K. I., Aonishi, K., & Kawasaki, Z. I. (2009). A Kalman filter approach to the Global Satellite Mapping of Precipitation (GSMaP) from combined passive microwave and infrared radiometric data. *Journal of the Meteorological Society of Japan. Series II, 87*, 137–151. https://doi.org/10.2151/jmsj.87a.137

Wu, H., Yang, Q., Liu, J., & Wang, G. (2020). A spatiotemporal deep fusion model for merging satellite and gauge precipitation in China. *Journal of Hydrology, 584*, 124664. https://doi.org/10.1016/j.jhydrol.2020.124664

Xie, P., & Xiong, A. Y. (2011). A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses. *Journal of Geophysical Research, 116*(D21), D21106. https://doi.org/10.1029/2011jd016118

Xu, R., Tian, F., Yang, L., Hu, H., Lu, H., & Hou, A. (2017). Ground validation of GPM IMERG and TRMM 3B42V7 rainfall products over southern Tibetan Plateau based on a high-density rain gauge network. *Journal of Geophysical Research: Atmospheres, 122*(2), 910–924. https://doi.org/10.1002/2016jd025418

Yang, K., & He, J. (2018). China meteorological forcing dataset (1979-2018). National Tibetan Plateau Data Center. https://doi.org/10.3972/westdc.002.2014.db

Yang, K., He, J., Tang, W., Qin, J., & Cheng, C. C. (2010). On downward shortwave and longwave radiations over high altitude regions: Observation and modeling in the Tibetan Plateau. *Agricultural and Forest Meteorology, 150*(1), 38–46. https://doi.org/10.1016/j.agrformet.2009.08.004

Zhang, T., Liu, Z. L., Wang, X. H., Xing, X. F., Chen, C. P., & Chen, E., (2018). Facial expression recognition via broad learning system. In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1898–1902). IEEE.

Zhao, G., Wang, X., Kong, Y., & Cheng, Y. (2021). Spectral-spatial Joint classification of hyperspectral image based on broad learning system. *Remote Sensing, 13*(4), 583. https://doi.org/10.3390/rs13040583

Zhao, H., Zheng, J., Xu, J., & Deng, W. (2019). Fault diagnosis method based on principal component analysis and broad learning system. *IEEE Access, 7*, 99263–99272. https://doi.org/10.1109/access.2019.2929094

Zhao, X., & Niu, R. (2019). Similarities and differences of summer persistent heavy rainfall and atmospheric circulation characteristics in the middle and lower reaches of the Yangtze River between 2016 and 1998. *Torrential Rain Disasters, 38*, 615–623.

Zhong, S. X. (2020). Diurnal variation of the duration and environment for heavy rainfall during the warm season in South China. *Atmospheric Science Letters, 21*(6), e969. https://doi.org/10.1002/asl.969

Ziarh, G. F., Shahid, S., Ismail, T. B., Asaduzzaman, M., & Dewan, A. (2021). Correcting bias of satellite rainfall data using physical empirical model. *Atmospheric Research, 251*, 105430. https://doi.org/10.1016/j.atmosres.2020.105430