Received: 6 November 2023 | Revised: 8 July 2024 | Accepted: 15 July 2024

DOI: 10.1002/gdj3.267

Geoscience Data Journal | RMetS | WILEY

DATA ARTICLE

Completion of the Central Italy daily precipitation instrumental data series from 1951 to 2019

Gamal AbdElNasser Allam Abouzied[1,2] | **Guoqiang Tang**[3] | **Simon Michael Papalexiou**[4] | **Martyn P. Clark**[5] | **Eleonora Aruffo**[6,7] | **Piero Di Carlo**[6,7]

[1]Department of Psychological Sciences, Health and Territory, University of the Studies "G. d'Annunzio", Chieti, Italy

[2]Drainage Research Institute (DRI), National Water Research Center (NWRC), Cairo, Egypt

[3]Climate and Global Dynamics, National Center for Atmospheric Research, Boulder, Colorado, USA

[4]Department of Civil Engineering, University of Calgary, Calgary, Alberta, Canada

[5]Coldwater Laboratory, University of Saskatchewan, Canmore, Alberta, Canada

[6]Department of Advanced Technologies in Medicine & Dentistry, University G. d'Annunzio, Chieti-Pescara, Italy

[7]Center for Advanced Studies and Technology– CAST, Chieti, Italy

Correspondence
Gamal AbdElNasser Allam Abouzied, Department of Psychological Sciences, Health and Territory, University of the Studies "G. d'Annunzio", Chieti, Italy.
Email: gamal.abouzied@unich.it

Funding information
Programma Nazionale FSE-FESR Ricerca e Innovzione 2014-2020 (PON) PhD fellowship, Grant/Award Number: DOT1753918

Abstract

Precipitation is a critical part of the global hydrological cycle that determines the distribution of water resources. It is also an essential meteorological variable used as input for hydroclimatic models and projections. However, precipitation data frequently lack complete series, especially at daily and sub-daily precipitation stations, which are usually large, bulky, and complex. To address this, gap filling is commonly used to produce complete hydrometeorological data series without missing values. Several gap-filling methods have been developed and improved. This study seeks to fill the gaps of 201 daily precipitation time series in Central Italy by localizing the approach used to generate the Serially Complete dataset for the Planet Earth (SC-Earth). This method combines the outcome of 15 strategies based on four various gap-filling techniques (quantile mapping, spatial interpolation, machine learning, and multi-strategy merging). These strategies employ the daily dataset of the neighbouring stations and the matched ERA5 data to estimate missing values at the target stations. Both raw data and the final serially complete station datasets (SCDs) underwent comprehensive quality control. Many accuracy indicators have been utilized to evaluate the performance of the strategies' estimations and the final SCD, such as Correlation Coefficient (CC), Root mean square error (RMSE), Relative bias (Bias %), and Kling-Gupta efficiency (KGE″). Multi-strategy merging strategy based on the Modified Kling-Gupta efficiency (MS_1) shows the highest performance as an individual precipitation gap-filling strategy. However, the machine learning strategy using random forest (ML_3) has the most outstanding share in the final estimates among all other strategies. In the end, the temporal–spatial performance of the final SCD is promising and depends on the pattern of the

The dataset is available at 10.5281/zenodo.12180685 and comprises five files. The CSV file contains essential station information, such as station ID, latitude, longitude, elevation, and station name. The observed, estimated, and final SCD are available in MAT format, with each column in these files representing a time series of a station in the same order as in the CSV file. Furthermore, the Nearby Info file includes the correlation coefficient (CC) and the distance between each target station and its assigned nearby stations, with each column corresponding to a target station in the same order as in the CSV file. The ERA5 data utilized in this study comprises hourly data on single levels from 1940 to the present (previously from 1950 to the present) sourced from the Copernicus Climate Data Store (CDS). The variable analysed is total precipitation (last access: 22nd April 2022).

missing values (MV%). The mean values of KGE″, CC, variability (α), and bias term (β) are 0.9, 0.93, 1.064, and 4.98×10^{-7}, respectively.

KEYWORDS

Central Italy, climate change, ERA5, gap filling, precipitation

1 | INTRODUCTION

The weather is the state of the lower atmosphere part, and individual meteorological elements, such as air temperature or rainfall, are the primary means to describe the weather. Although not all aspects are equally crucial to different population segments, the weather significantly impacts society (Nonhebel, 1993). Rainfall is one of the essential inputs in multiple disciplines, such as climatology, meteorology, irrigation engineering, irrigation scheduling, hydrology, water resource management, and environmental hazard assessment (floods) (Chinasho et al., 2021).

The majority of instrumental time series experience a percentage of missing values (MV%). Data gaps extend beyond the early instrumental era, where records were often lost due to factors like wars or fire. They also occur in the contemporary period, thanks to sporadic station interruptions, instrument malfunctions, and network reorganizations, among other reasons (Qin et al., 2021). The ability to model a wide range of hydrological and environmental processes requires the availability of a trustworthy source of rainfall and other climate data. While the form and structure of hydrological and environmental models differ, they always need a complete and reliable precipitation dataset on a temporal and spatial basis (Larson & Peck, 1974; Vieux, 2001). Unfortunately, the existence of gaps in meteorological time series denotes a widespread but critical issue since it can produce biased results. In the worst case, it can even prevent fundamental analyses of the considered variables from being carried out (Lo Presti et al., 2010). Also, these temporal discontinuities, besides poor data quality and inadequate observation periods, lead to a reduction in the number of available stations to be used in model analysis (Eischeid et al., 2000).

Many climatologic analyses are concerned with the treatment of missing data in meteorological time series. These analyses often focus on studies of droughts and above/below-threshold events, such as those based on well-known indices developed by the World Meteorological Organization Commission for Climatology/World Climate Research Program project on Climate Variability and Predictability and the Expert Team on Climate Change Detection, Monitoring, and Indices (Alexander et al., 2006). Excluding missing values from data analysis or ignoring the problem if the amount is insignificant is one technique to get around this challenge. However, such methodologies may overlook important information and biases in many climate studies (Simolo et al., 2010). To overcome this problem, several techniques have been developed over the last decades to estimate missing values in climatic time series on a seasonal, monthly, and daily basis. Thus, achieving accurate estimations of missing data in daily precipitation records remains challenging even if long-time series and intensive rain-gauge networks are considered (Simolo et al., 2010).

Multiple methods exist for addressing missing values. Nevertheless, among the most commonly utilized techniques are the following four:

1. Within-station methods are basic techniques for estimating missing data points in climate records. These methods are quite straightforward because they only rely on information from the same weather time series where the missing data occurred. For example, using the neighbouring days, this approach takes the average of the values recorded on the days just before and after the missing day (e.g., Kemp et al., 1983). However, it assumes that the weather conditions on a missing day would be similar to the surrounding days. These within-station methods are simple to apply, but they have limitations, such as they work well for climate variables with a strong similarity pattern between nearby time points (high autocorrelation). For example, temperature data often have such patterns. Also, they are suitable for calculating long-term averages because they rely on overall trends in the data. Nevertheless, they could be more beneficial in estimating the variables that can vary significantly from day to day and only sometimes follow a consistent pattern, so these basic methods may not accurately estimate daily rainfall amounts.
2. Spatial interpolation methods (INT) are commonly employed to fill gaps in weather time series, specifically precipitation time series, whether monthly or daily. These techniques rely on data collected simultaneously from nearby weather stations to estimate values for the missing data points at a target station. Various spatial interpolation methods are available, with one familiar approach being the inverse distance weighting method (Cressman, 1959; Shepard, 1968). In this method, the

proximity of the target station to surrounding stations is considered, and a weighted average is computed using the inverse of the squared distances as weighting factors. This approach assumes that nearby weather stations have a positive spatial correlation in their precipitation data. However, using distance alone as a similarity criterion for precipitation time series can be inadequate, and the selection of nearby stations is crucial for the accuracy of the results. Several modifications and extensions to the inverse distance weighting method have been proposed to address these limitations. These include using higher powers of distance, negative exponential functions of distance (Garcia-Marin et al., 2008; Teegavarapu & Chandramouli, 2005), or considering topographical factors like orographic effects instead of the inverse squared distance (Daly et al., 1994; Lloyd, 2005). In addition to inverse distance weighting, other spatial interpolation methods commonly used for daily data include spline-surface fitting, optimal interpolation, multi-linear regression (MLR), and kriging (Eischeid et al., 2000). Depending on the specific data and geographic context, these techniques help estimate missing values with varying complexity and accuracy.

3. The quantile mapping (QM) method operates on the hypothesis that the estimated data distribution mirrors the observed data distribution. Within QM, estimated data linked to a specific probability is exchanged with an observed quantile that aligns with the same probability. The choice of probability distribution models for both observed and estimated data is pivotal to the QM process. Therefore, carefully selecting a fitting probability distribution model is of utmost importance for successfully applying the QM method (Heo et al., 2019). The empirical cumulative distribution function (CDF) is commonly employed to describe the probability distribution of precipitation as it effectively reduces inaccuracies by preserving the frequency distribution of modelled and observed precipitation data (Prudhomme et al., 2012). It is particularly useful in preventing an overestimation of rainy days in precipitation records while maintaining the distribution of time series, which is valuable for estimating extreme events (Cannon et al., 2015). Grillakis et al., 2020 demonstrated the adequate performance of empirical quantile mapping in filling the missing daily rainfall data on the Mediterranean island of Crete.
4. Machine learning, as a data-driven approach, stands out as an effective option for addressing missing data. Some studies have introduced machine learning techniques like artificial neural networks (ANN), support vector machines (SVM), and random forests (RF) to address gaps in precipitation data. Artificial neural networks, for instance, have the capability to derive patterns and relationships from a limited subset of data while also sustaining robustness in the existence of noise or missing data (Ilunga & Stephenson, 2005). Additionally, Long-Short Term Memory (LSTM), a powerful architecture within the realm of artificial recurrent neural networks (RNN), has proven to be highly effective in various applications, recently demonstrating strong performance in filling data gaps.

Numerous studies in the context of time series gap-filling have been conducted at various scales. Local studies, such as Nkuna and Odiyo (2011) in the area of Luvuvhu River Catchment, Mwale et al. (2012) in the Shire River basin, Malawi, Aguilera et al. (2020) in SW Spain, Chinasho et al. (2021) in South Ethiopia, Bellido-Jiménez et al. (2021) in Andalusia – Southern Spain, Simolo et al. (2010) in Reno River basin in northern Italy, and Di Piazza et al. (2011) in Sicily, Italy. However, only a limited number of studies have been conducted on a global scale. Tang et al. (2021) developed the Serially Complete Earth (SC-Earth) dataset, which provides daily precipitation, mean temperature, temperature range, dewpoint temperature, and wind speed data from 1951 to 2019. Additionally, a study covering North America resulted in the provision of serially complete datasets (SCD) for precipitation and minimum and maximum temperature data from 1979 to 2019 (Tang et al., 2020). Nevertheless, the study of global precipitation encounters significant obstacles owing to its extensive spatial and temporal fluctuations, which exceed those observed in temperature. Conversely, the reliability of precipitation data is compromised by the diverse array of measurement equipment and methodologies utilized across distinct countries and regions.

For the previously mentioned reasons, this research aims to generate a complete daily rainfall dataset for Central Italy spanning from 1951 to 2019, which will be achieved through regional downscaling of the SC-Earth approach (Tang et al., 2021) that involves various types of gap-filling methods. The resulting dataset can reduce the error level as it consists of an extensive and well-distributed network. Simultaneously, it can support the researchers and policymakers in the country, offering valuable insights into climate variability and change in one of the most topographically complex regions.

2 | STUDY AREA AND MATERIAL

2.1 | Station data

The study area (Figure 1b) spans practically two regions in central Italy toward the south (the entire region of Abruzzo, Molise, and small parts of Apulia, Lazio, and Campania),

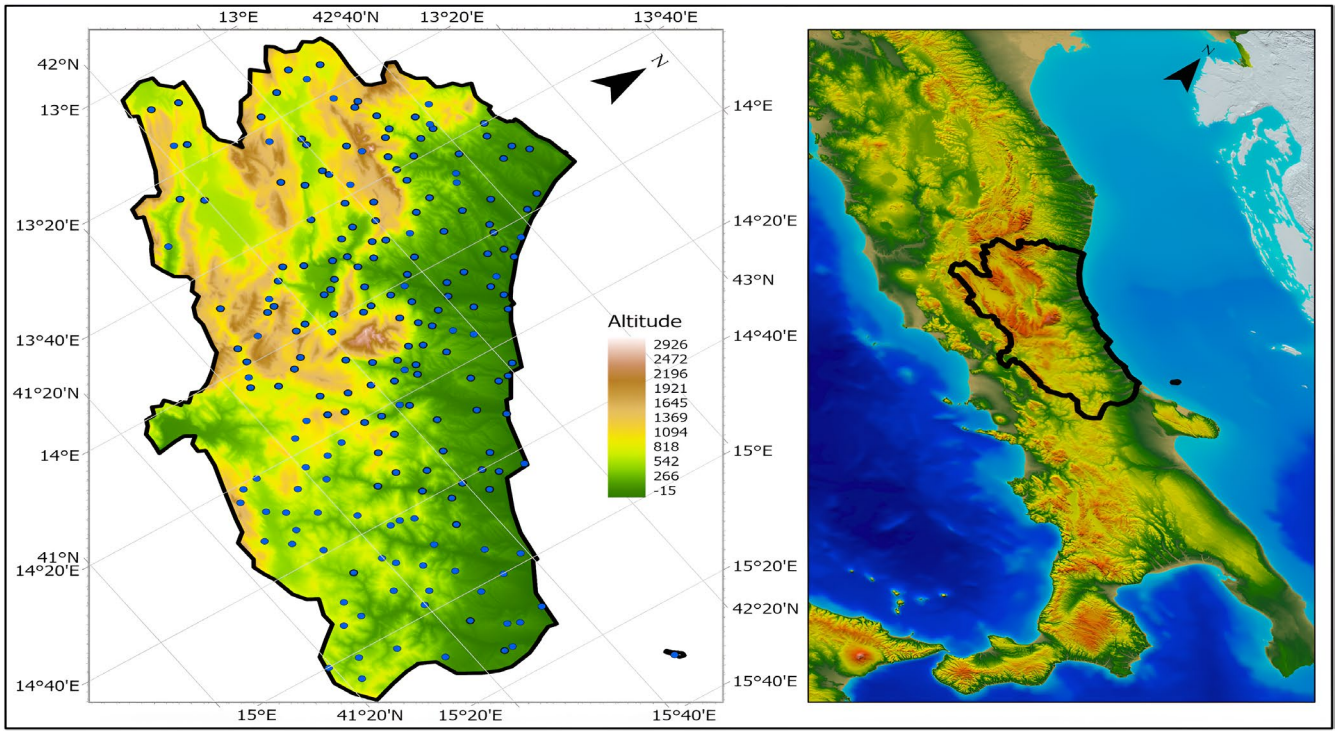

FIGURE 1 (a) Localization of the study area in Central Italy, (b) digital elevation model for the study area to show the area's topography and the locations of the selected stations.

with latitudes ranging from 41°19′ to 42°50′ N and longitudes ranging from 13°07′ to 14°45′ E. The diverse topography, with heights varying from sea level to the highest peak of the Apennine (2912 m), significantly impacts the surrounding climate. The Apennine Mountains, in fact, create more intense precipitation with height due to the Stau effect (i.e., orographic lift). At the same time, due to the protection provided by these mountain ranges against Adriatic and Tyrrhenian low-pressure systems, the inland valleys receive the least amount of annual rainfall (i.e., rain shadow) (Leopardi & Scorzini, 2015).

Rainfall data for this study come from an extensive monitoring network (Figure 1a) operated by the regional public Hydrographic and Mareographic Office (Ufficio Idrografico e Mareografico, Regione Abruzzo, https://www.regione.abruzzo.it/content/idrografico-mareografico). The original precipitation time series comprises daily cumulated rainfall measured at 284 meteorological sites. Many of these gauges have provided monthly time series since the second half of the nineteenth century. Yet, daily data are only available since the second half of the twentieth century.

The selection of the input dataset was based on criteria that emphasized period covered, completeness, and spatial coverage as much as possible. Stations whose time series with an extremely large percentage of missing values (MV%), greater than 92% for the study period from 1951 to 2019, and low monthly or seasonal samples are excluded. As a result, 205 stations were selected out of the 284. None of these 205 stations has a complete time series. Two stations have MV% of less than 1%, whereas 39% have MV% of more than 50%, such as those located in the southern part of the study area, and 5% have high missingness, with an MV% of more than 80% (Figure 2a). Also, the figure illustrates that stations in the neighbouring regions exhibit elevated MV% values compared to Abruzzo. This discrepancy could be attributed to the data provider's capacity to gather data from outside Abruzzo.

The temporal distribution of the MV% (Figure 2a,b) is divided into three sub-periods because of the various data resources and formats. From 1951 to 1973, the MV% was almost stable at around 37%, then a sharp drop in MV% in 1973 until the minimum value (12.44%) before MV% started a slight increment until 2006. After that, a significant increase in MV% reached the maximum value of 93.75% at the end of the study period. As per the hydrographic office, the increase in the MV% in recent years is attributed to a lack of funds, resulting in insufficient maintenance and the closure of some stations.

Many gaps and extended durations of missing time intervals, (presented in Figure 2c), characterized most of the time series. There are 1826 gaps in the selected time series. The lengths of these gaps vary, ranging from 1 month to 56 years days; 45% of the gaps are a month-long, while 34.5% are between a month and a year-long, and the rest are from over a year-long to 56 years.

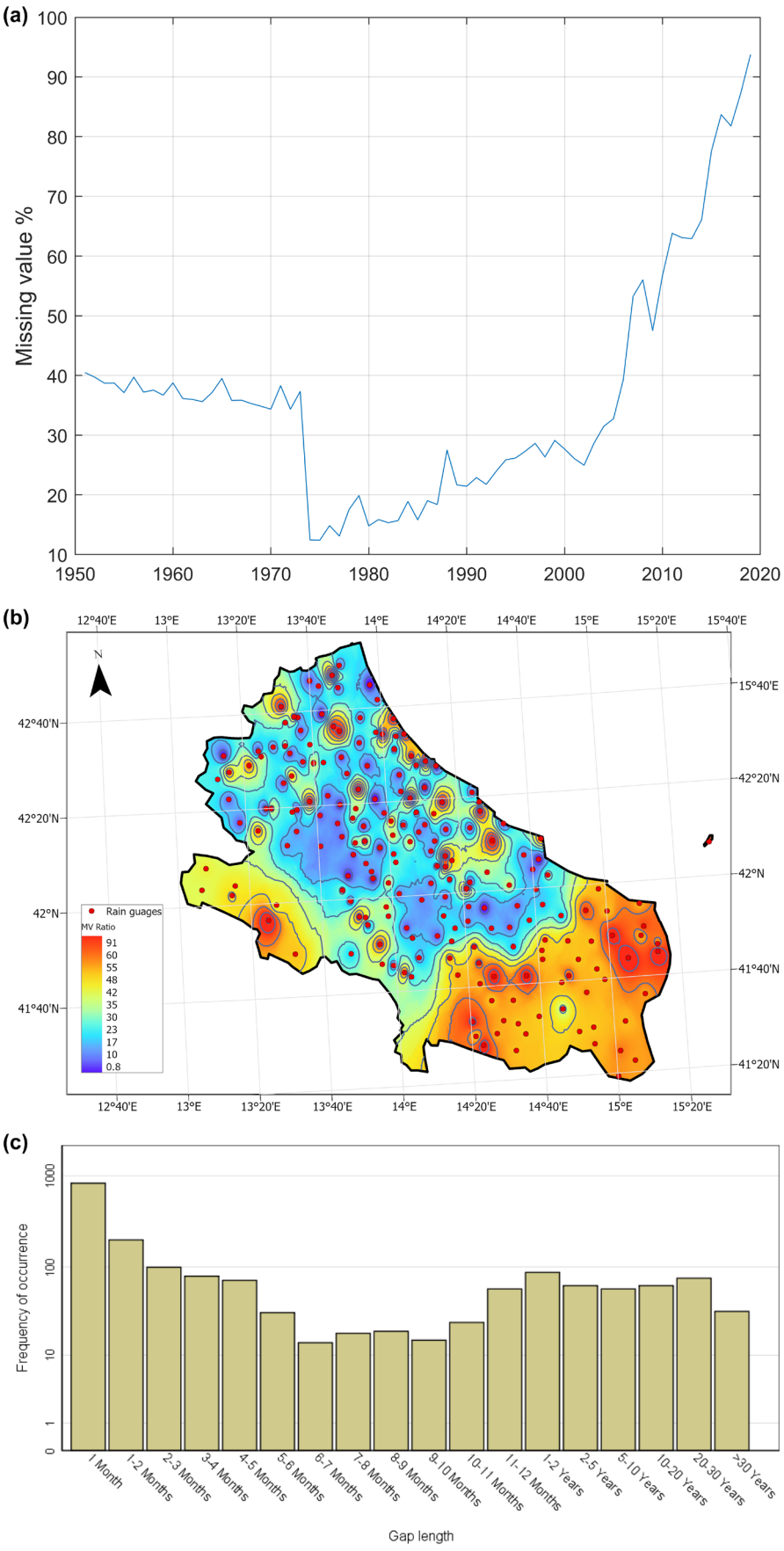

FIGURE 2 (a) Temporal variation of missing value ratio, (b) spatial distribution of the missing value ratio, and (c) lengths of gaps and their occurrence frequency.

2.2 | ERA5 data

ERA5 is a new climate global reanalysis dataset built by the European Centre for Medium-Range Weather Forecasts (ECMWF) in March 2019. ERA5 has several significant advancements compared to previous ECMWF Reanalysis ERA-Interim, one of the best-performing products in hydrological studies, an increase in the horizontal grid spacing, the number of vertical levels, the temporal resolution, and the number of observations assimilated (Beck, van Dijk, et al., 2017; Beck, Vergopolan, et al., 2017; Tarek et al., 2020). ERA5 assessments are based on the Integrated Forecasting System (IFS) Cycle 41r2. This reanalysis integrates new input variables, like sea surface temperature, sea ice, and aerosols, seeking to make it suitable for climate simulation. The ERA5 data is available for free on the website of ECMWF (ECMWF, 2017), with rainfall estimations for the period from 1979 onwards. However, for some meteorological elements, the dataset provides information for the period since 1940 (Quagraine et al., 2020; Tarek et al., 2020). One hour and $0.25° \times 0.25°$ are the maximum temporal and spatial resolutions available on ERA5 (Hersbach et al., 2020).

3 | METHODOLOGY

3.1 | Data pre-analysis

Data preparation encompasses a set of procedures focused on scrutinizing raw data to generate high-quality data. These procedures primarily encompass data collection, integration, conversion, cleaning, removal, and discretization. In this study, data retrieval involved extracting information from various resource files, such as xlsx and txt files, covering distinct sections of the study area and different durations within the study period.

Subsequently, the data underwent a format validation process to identify potential issues, including the absence of months or days, invalid characters in data fields, and negative values. During this phase, any records found to be problematic were identified as missing. Following this, the results were aggregated to create daily time series datasets for the available stations within the study area.

To further refine the dataset, a straightforward filter was applied based on the completeness of the time series data and the availability of spatial data. This filter facilitated selecting the most appropriate time series for the subsequent gap-filling procedure.

3.2 | Quality control

Behind the stations' data preparation, a strict and comprehensive quality control approach was applied. Initially, this approach was used to build SC-Earth (Tang et al., 2021) and SCDNA (Tang et al., 2020). This method consists of three parts:

1. The first part applies different checks (Durre et al., 2010): integrity checks, outlier checks, internal and temporal consistency checks, spatial consistency checks, and extreme mega consistency checks.
2. The second phase of quality control (QC), as outlined by (Hamada et al., 2011), closely resembles the initial phase and encompasses the following set of procedures: repetition checks, duplicated monthly or sub-monthly record check, *Z*-score-based outlier check, and spatiotemporally isolated value check.
3. The third segment comprises two types of assessments introduced by Beck et al. (2019). It is utilized to examine the percentage of days without precipitation and validate unique values.

Values that do not meet quality control standards are regarded as missing.

3.3 | Construct the SCD

Numerous research findings illustrate that enhanced performance can be achieved by combining various infilling and reconstruction techniques rather than relying on individual methods. In this study, the combined infilling gap techniques of SC Earth (Tang et al., 2021) and SCDNA (Tang et al., 2020) are employed, which incorporate the following distinct strategies:

1. Four of them rely on quantile mapping with nearby stations, e.g., (Longman et al., 2019; Newman et al., 2015):
   QM_1: Closest nearby stations
   QM_2: Weight-mean using Correlation Coefficient.
   QM_3: Weight-mean using distance.
   QM_4: Median of the previous three.
2. One is based on quantile mapping with concurrent ERA5 estimations (QMR).
3. Four are based on spatial interpolation methods:
   INT_1: MLAD interpolation.
   INT_2: Normal Ratio interpolation.
   INT_3: IDW interpolation.
   INT_4: Median of the previous three.

4. Four are based on machine learning techniques:
   ML_1: Artificial Neural Networks (ANN).
   ML_2: Random Forest model (RF).
   Ml_3: Long-Short Term Memory (LSTM).
   ML_4: Median of ANN/RF/LSTM
5. Two use multi-strategy merging (MRG).
   MS_1: Rank Modified Kling-Gupta efficiency of (QM_1 to QM_3, QMR, INT_1 to INT_3, and ML1 to ML_3).
   MS_2: The median of the three selected strategies in MS_1.

For each candidate station and day in the study period, the adapted technique comprises the nine processes listed below:

Step 1: For each station, extract the spatiotemporally concurrent reanalysis estimates (ERA5) and match the value of ERA5 estimations with the actual station observations.

Step 2: Defining the nearby stations according to the overlapping period (at least 8 years) and the distance (<200 km). The nearby stations should be at least one and no more than 30. These stations are arranged in order of their correlation with the target station.

Step 3: Using a 31-day time window centred by the target day to determine the empirical cumulative distribution functions (CDFs) of the target station, neighbouring stations and ERA5 estimations.

Step 4: Using the previously mentioned strategies to generate the estimated daily values of the study period for the target station.

Step 5: For validation objectives, repeating the last two steps (steps 3 and 4) using 70% of the actual measurements, and the remaining 30% of observations are utilized to validate the 15 strategies independently.

Step 6: Ordering the strategies' estimations (step 4 outputs) according to their accuracy index derived from the last step (step 5), the highest accuracy strategy will be adopted.

Step 7: Correcting the selected estimations to be close to the actual measurements as possible by applying wet-day bias on the estimated values (Beck et al., 2019) and mean value correction, which is rescaling the estimated series by using the ratio between the calculated mean values of estimations and observations for the observed period.

Step 8: To generate the final SCD, the estimates should be replaced by the actual measurements whenever it is possible.

Step 9: Running a quality control check on the final SCD.

3.4 | Performance evaluation

The temporal–spatial performance of ERA5 precipitation data, gap-filling methods estimations, and the final SCD was assessed in this study using one or some of the following accuracy indicators:

3.4.1 | Pearson's linear correlation coefficient (CC)

The strength of the relationship between any two-time series, such as the observed and estimated or the ERA5 precipitation for a particular station, can be quantified from the Pearson correlation coefficient. It also illustrates their linear association and demonstrates the method's suitability for forecasting missing data. The CC values vary from 0 to 1, and the greater the CC value, the more likely it is that the two time series follow the same tendency.

$$CC = \frac{\sum_{i=1}^{n}\left(X_e - X_e^*\right)\left(X_o - X_o^*\right)}{\sqrt{\sum_{i=1}^{n}\left(X_e - X_e^*\right)^2\left(X_o - X_o^*\right)^2}} \quad (1)$$

where X is the precipitation value, X^* is the mean value, the subscripts e and o refer to estimations or ERA5 reanalysed data and observed data, n is the length of the time series, and i is the day number.

3.4.2 | Root mean square error (RMSE)

The RMSE is typically used to assess the effectiveness and performance of the different estimated methods in meteorological studies. The degree of overall error between the estimated or ERA5 reanalysed and observed data is represented by RMSE. The range of the RMSE values is 0 mm/day to $+\infty$ mm/day, and the most optimistic estimates data are that with the lowest RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_e - X_o\right)^2}{n}} \quad (2)$$

3.4.3 | Relative bias (Bias %)

The bias% value signifies the degree of deviation between the precipitation data derived from gap-filling methods or ERA5 and the actual observed precipitation data. Bias% falls within the range of −100% to +100%, with the optimal value being 0%.

$$Bias\,\% = \frac{\sum_{i=1}^{n}\left(X_e - X_o\right) \times 100}{n \times X_o^*} \quad (3)$$

3.4.4 | Modified Kling-Gupta efficiency (KGE″)

Efficiency criteria are commonly used to assess deterministic simulations' accuracy. Gupta et al. (2009) introduced the Kling-Gupta efficiency (KGE) criterion, which is a new criterion that was later improved into the modified criteria known as KGE′ and KGE″ by Kling et al. (2012) and Santos et al. (2018) consecutively. KGE″ integrates

the correlation, bias, ratio of variances or coefficients of variation in a more balanced technique. Moreover, it prevents abnormally negative KGE or KGE′ results when the mean value is close to zero. In addition to evaluating and ordering the strategies' estimations (step 6), KGE″ is used to facilitate the merging of several techniques and the evaluation of the estimated precipitation. The KGE″ value's range is 0 to +1 and the higher KGE″, the more efficient estimates:

$$\mathrm{KGE}'' = 1 - \sqrt{(\mathrm{CC}-1)^2 + (\alpha-1)^2 + \beta^2} \quad (4)$$

where CC is Pearson's linear correlation coefficient between observations and estimates, $\alpha = \sigma_e / \sigma_o$ is the variability ratio, σ_o and σ_e are the standard deviations of observations and estimates, $\beta = (\mu_e - \mu_o) / \sigma_o$ is the bias term, and μ_o and μ_e are the mean values of observations and estimates.

4 | RESULTS AND DISCUSSIONS

4.1 | Quality control

Only two out of the initial 205 stations with daily precipitation data failed to pass the first quality control test. A hundred forty-one stations passed this test without losing a single value, and the remaining 62 stations had at least one daily value and a maximum of 2 years removed during this test. When comparing the temporal availability before and after the quality control, it was discovered that roughly 2.6% of the data from 2005 to 2019 had poor quality (Figure 3). In total, 0.6% of the daily dataset was rejected because of quality issues.

4.2 | Gap-filling strategies

4.2.1 | Evaluation of ERA5

This study utilizes daily ERA5 precipitation data on standard latitude-longitude grids at 0.25° × 0.25° from 1950 to 2019 to extract ERA5 rainfall data at the stations' locations. The CC, bias%, and RMSE are the statistical validation indicators used to assess the temporal–spatial performance of ERA5 precipitation data.

As illustrated in Figure 4, it is evident that the daily ERA5 precipitation exhibits relatively weak correlations with the observed data, both in spatial and temporal dimensions, with the mean CC value of 0.42 in both. The RMSE reflects the general level of error between the ERA5 and the observed precipitation data, which can represent data stability. The RMSE, which can represent data stability, has a mean value of 6.6 mm/day for the temporal

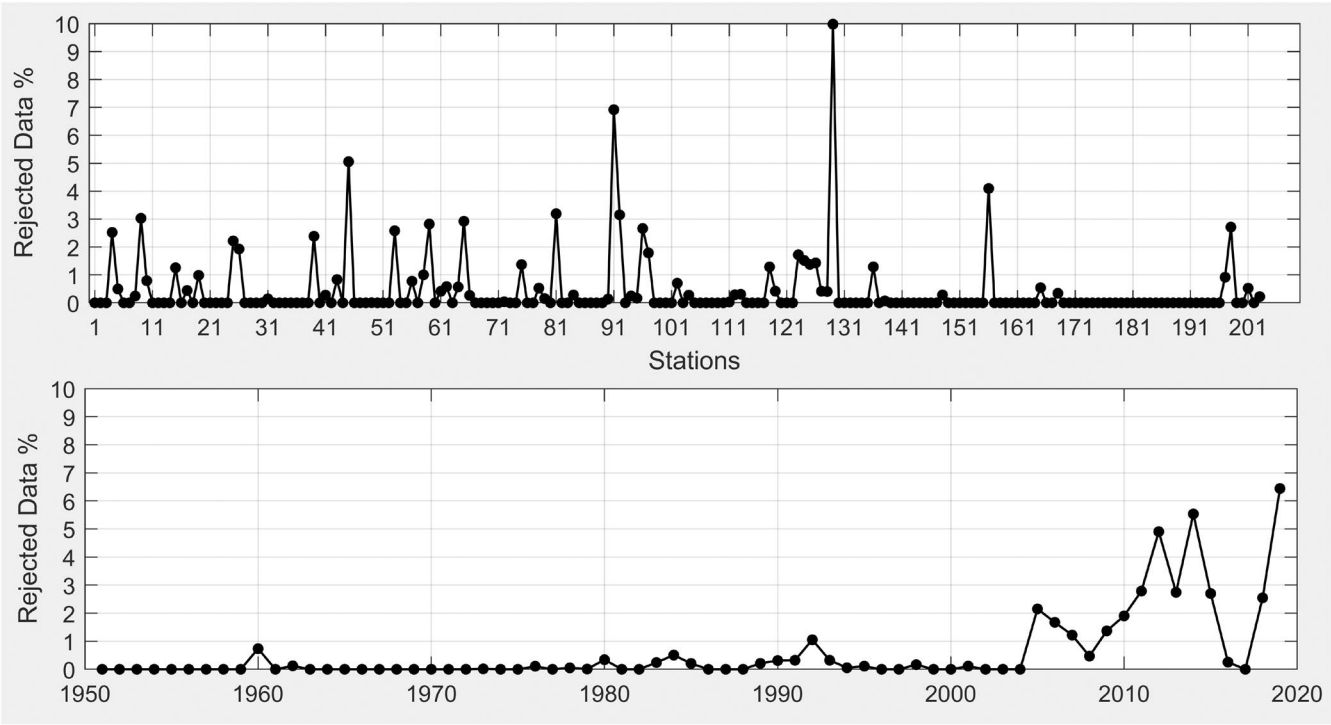

FIGURE 3 Percentage of daily precipitation data that failed in QC relative to observed data across stations (upper panel) and time (lower panel). As shown, a lot of values have been rejected in recent years because of the contributions from the two removed stations. Also, many stations lost their funding, and there was a reduction in calibration and maintenance.

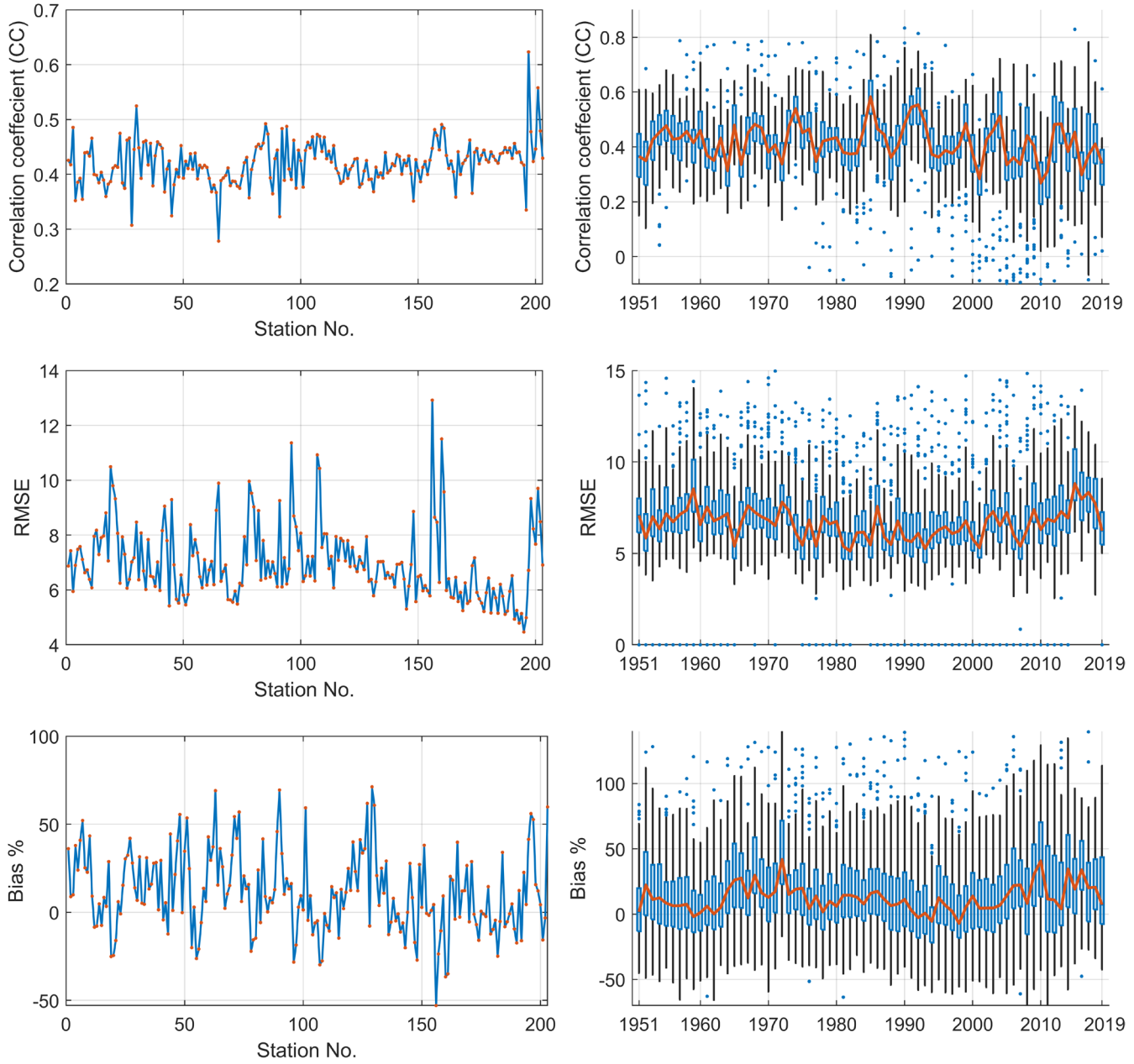

FIGURE 4 The CC, bias%, and RMSE (mm/day) between the ERA5 and observed data. The left panel is plotted according to the spatial dimension (station) while the right panel is plotted according to the temporal dimension and the red line represents the median values.

dimension and 6.99 mm/day for the spatial dimension, indicating that the ERA5 data has a marginally low error amount. The temporal and spatial bias% mean values are 12.15% and 11.50%, respectively, which signifies that the ERA5 data has a low deviation from the observation data.

4.2.2 | Selection of neighbouring stations

Due to the dense station network within the relatively small study area, each station enjoys the maximum number of neighbours, totalling 30 stations. The average daily precipitation correlation between pairs of observatories with a minimum of 8 years of shared data shows a notable decrease in correlation as a function of distance, ranging from 1 to 110 km (mean CC drops from 0.85 to 0.53), as illustrated in Figure 5. This decline remains relatively modest but consistent at greater distances.

In general, the correlation coefficient (CC) between the target station and its neighbouring stations is significantly higher than that between the target station and the reanalysed ERA5 data. This observation underscores the higher reliability of neighbouring stations as a data source for the gap-filling process.

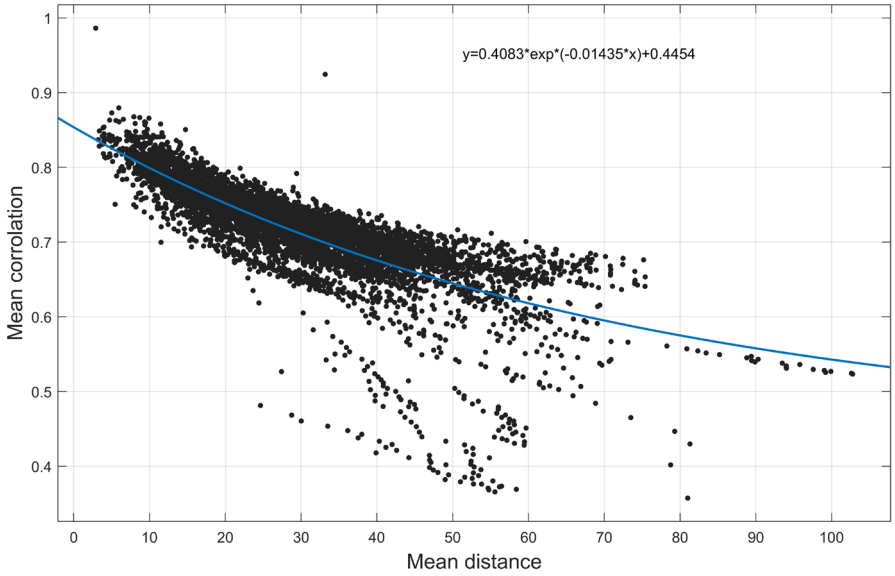

FIGURE 5 The mean CC and the mean distance between the daily precipitation series and their neighbourhood stations for all possible pairs. Additionally, the graphic displays the best-fitting curve and its equation.

TABLE 1 Medians and means of RSME (mm/day), CC, and KGE″ calculated for each proposed strategy.

	RMSE (mm/day)		CC		KGE″	
Strategy	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
QM_1	4.49	4.65	0.78	0.79	0.79	0.78
QM_2	3.59	3.82	0.86	0.84	0.82	0.81
QM_3	3.67	3.86	0.85	0.84	0.83	0.82
QM_4	3.61	3.79	0.86	0.85	0.83	0.82
QMR	5.58	5.87	0.66	0.65	0.65	0.65
INT_1	3.46	3.64	0.87	0.86	0.80	0.78
INT_2	3.77	3.94	0.84	0.83	0.81	0.80
INT_3	3.76	3.95	0.84	0.83	0.81	0.80
INT_4	3.51	3.71	0.86	0.85	0.81	0.80
ML_1	3.95	4.17	0.82	0.81	0.79	0.77
ML_2	3.01	3.24	0.90	0.89	0.84	0.82
ML_3	2.93	3.15	0.90	0.89	0.89	0.87
ML_4	2.93	3.16	0.90	0.89	0.83	0.82
MS_1	2.37	2.57	0.94	0.93	0.90	0.89
MS_2	2.68	2.84	0.92	0.91	0.88	0.87

4.2.3 | Performance evaluation of candidate gap-filling strategies

The accuracy of estimating precipitation occurrences is assessed using three performance metrics: RMSE, CC, and KGE. These metrics are applied to evaluate the precipitation estimates obtained from the proposed techniques in comparison to the observed data from rain gauges (as shown in Table 1, Figure 6).

The RMSE findings indicate that machine learning and multi-strategy merging exhibit lower median RMSE values in comparison to other methods. Specifically, MS_1 demonstrates the lowest RMSE at 2.37 mm/day, while QMR exhibits the highest RMSE at 5.58 mm/day.

Nevertheless, a noteworthy correlation exists between the estimated and observed data, except in the case of the QMR strategy, which exhibits a weaker correlation compared to other approaches. The findings reveal that the median value of CC surpasses 0.8 in all the proposed methodologies, except for those mentioned above. The multi-strategy merging approach boasts the highest median CC (ranging from 0.938 to 0.92) for the two strategies considered.

The KGE is regarded as the accuracy meter and the principal performance measure. The median KGE values for the proposed approaches exhibit a similar trend to the correlation. With the exception of the QMR method, all strategies attain commendable KGE values. In terms of the spatial performance of this crucial meter (Figure 7),

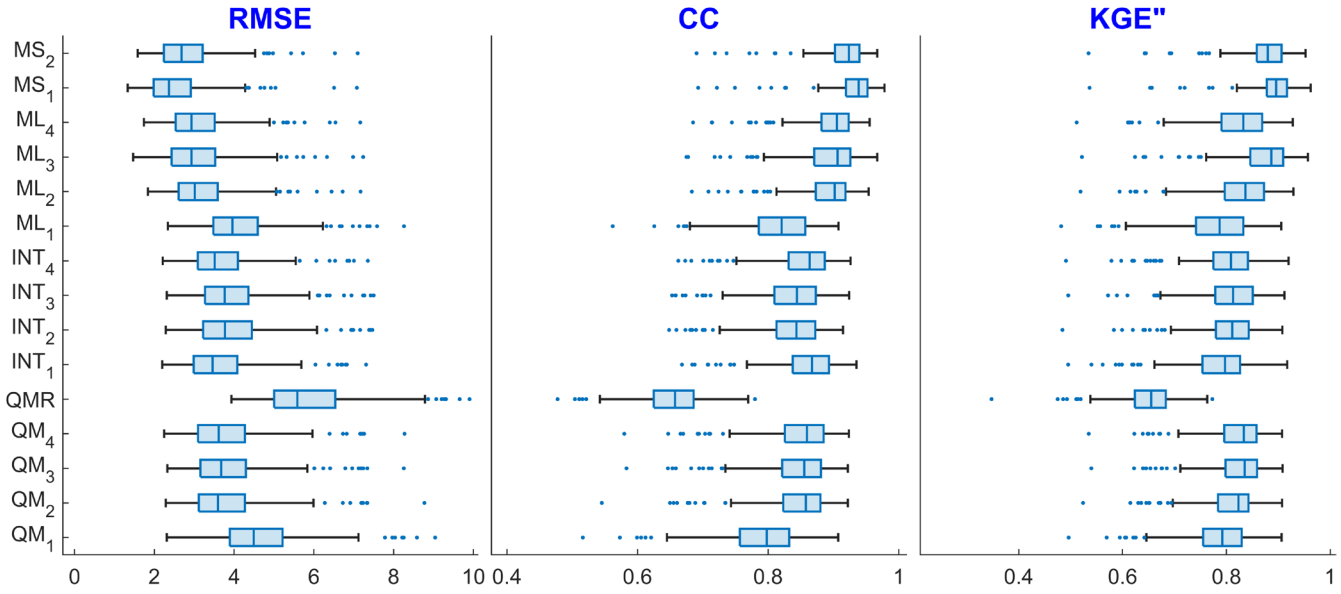

FIGURE 6 RSME (mm/day), CC, and the KGE″ calculated for each station and each proposed strategy.

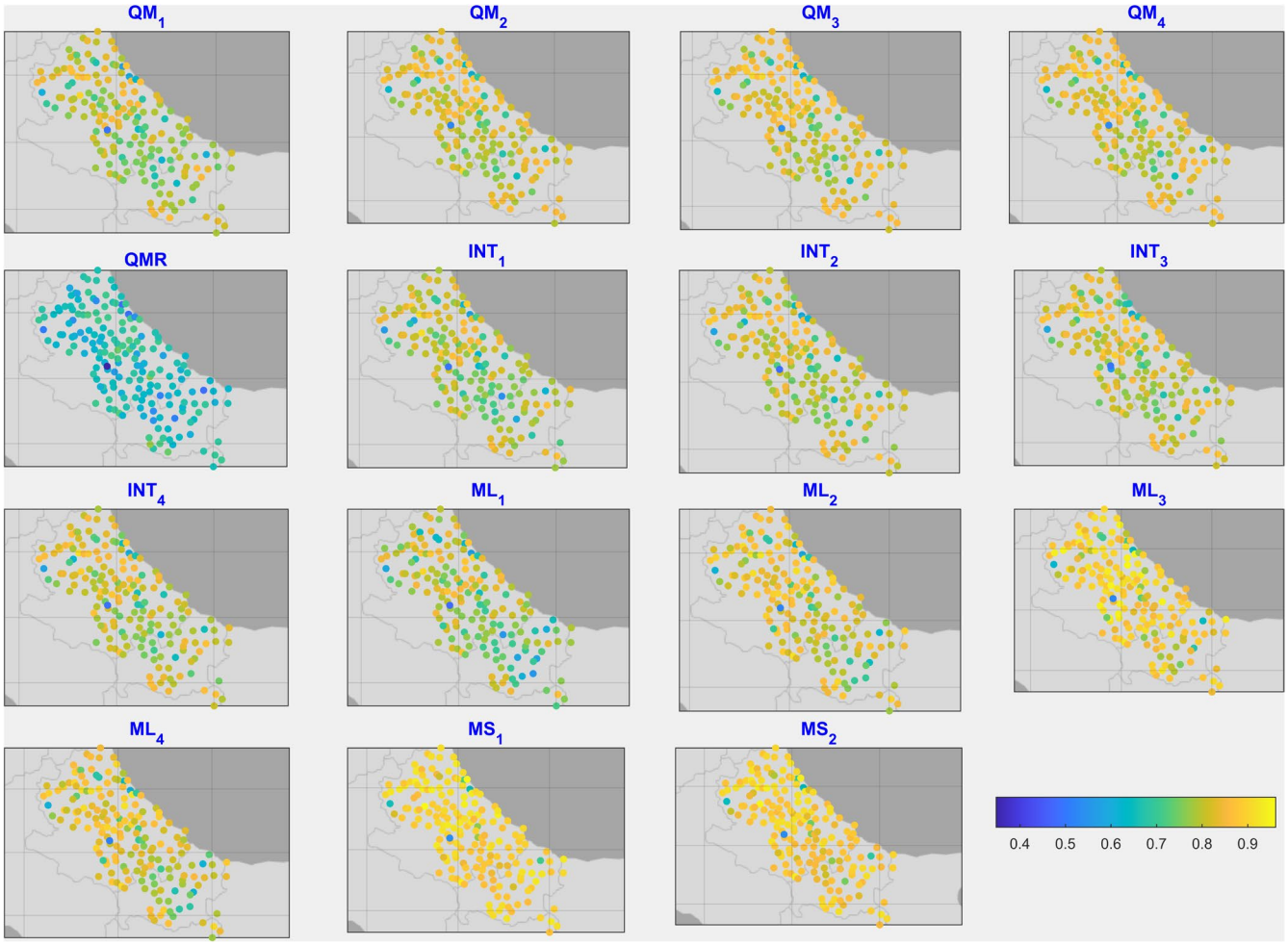

FIGURE 7 The spatial distributions of KGE″ for precipitation estimates from 15 gap-filling strategies.

all strategies exhibited nearly identical performance across the entire study area despite the intricate topography. Nevertheless, the high station density in the northern part of the region can pose an improvement to the effectiveness of gap-filling due to this concentration.

In general, the performance of the INT_1 and ML_1 techniques tends to be less satisfactory, likely owing to their extensive demand for training samples in both multiple linear regression and artificial neural networks (Tang et al., 2021). Conversely, the QMR technique performs poorly, indicating that ERA5 may not offer sufficient precision for gap filling, even when employing the best temporal matching.

On the contrary, the multi-strategy merging approach surpasses all other proposed strategies, boasting the highest median values for CC and KGE″ and the lowest RSME, especially in the case of MS_1. Overall, LSTM stands out as the most effective individual method for filling precipitation gaps, consistent with findings reported by Tang et al., 2021. This outstanding performance is attributed to the ability of LSTM and other deep-learning models to address the nonlinear and periodic challenges associated with rainfall prediction.

4.3 | The estimated serially complete dataset (SCD)

The Serially Complete Station Dataset (SCD) is generated by sequentially aggregating the outputs of all 15 gap-filling techniques. The Machine Learning (ML) group plays a predominant role, contributing approximately half of the total estimated dataset. Within the ML group, there are significant variations in the contribution percentages of individual techniques, with ML3 leading the way at 35%, marking it as the most influential strategy among all others, not just within the ML group. This aligns with the previously established notion that ML3 stands out as the most effective individual strategy.

The Multi-Strategy (MS) group follows closely with the second-highest percentage of participation at 32%, with no noticeable disparity between the two MS approaches. The remaining 18% is evenly distributed among the QM, QMR, and INT groups, with QMR claiming the smallest share at less than 1%, as depicted in Figure 8.

Furthermore, there exists a nearly linear relationship between the values of the SCD estimations and the observations. This underscores the high quality of the SCD estimations and the absence of statistically significant differences between the observed and estimated data (Figure 9).

4.4 | Performance evaluation of the final SCD

The final Serially Complete Station Dataset (SCD) was created by merging the estimated SCD with the observed data, following a correction and rescaling procedure applied to the estimated SCD. The critical advantage of SCDs over pure spatial interpolation lies in their utilization of observational data from the target station to constrain the SCD estimates. This attribute accounts for the remarkable alignment between the final SCD estimates and the actual observations (Tang et al., 2021).

Figure 10 presents the monthly data for six stations as an example, although the study encompasses 201 sites in total. These selected stations exhibit varying degrees of missing data percentages (ranging from 0.73% to 86.96%) and gap sizes, effectively representing the broader spectrum of stations in the study. The resultant SCD estimates effectively eliminate gaps throughout the entire time series. However, the robust connection between the final

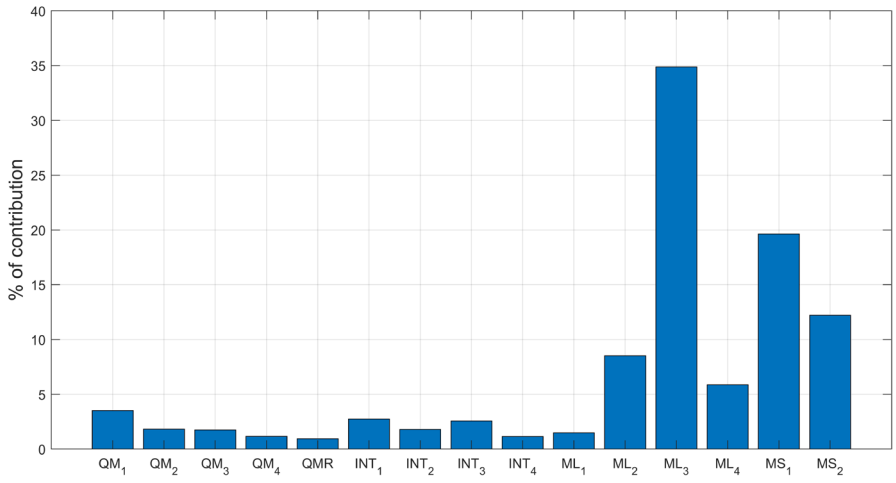

FIGURE 8 The 15 strategies' percentage of contribution in the final serially complete datasets.

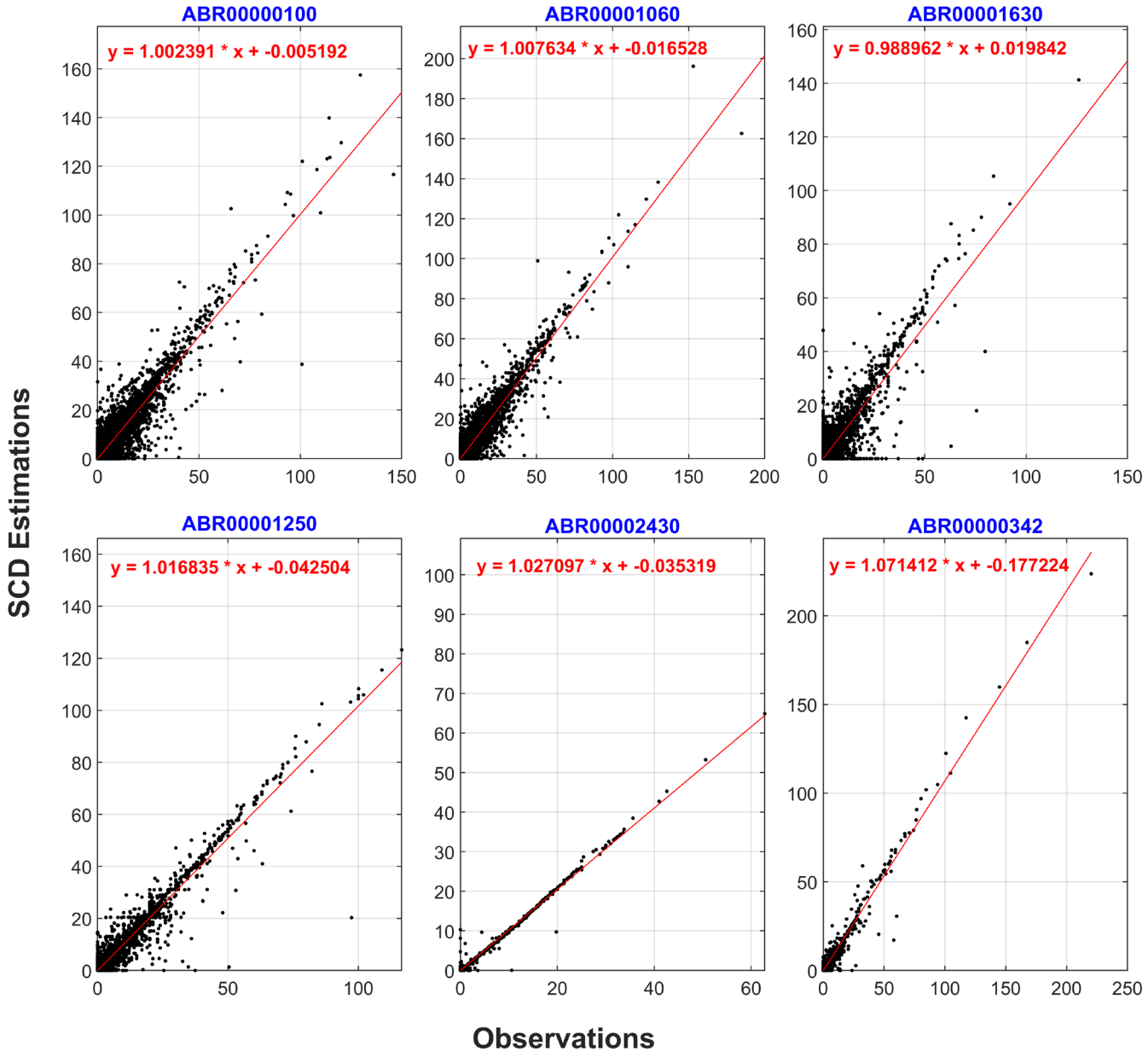

FIGURE 9 Comparing daily data between SCD estimations and observations, along with the inclusion of a best-fit line, is demonstrated using six example stations that exhibit diverse characteristics. These stations are representative samples of the entire set.

SCD and the observed data is more noticeable in the daily data curves of these stations.

4.4.1 | Spatial distribution of the final SCD's accuracy

Considering the spatial allocation of KGE″ for the final SCD and its three components (CC, α, and β), it's evident that CC exhibits similar distributions to KGE″, with respective mean values of 0.9 and 0.93. These values underscore the remarkable accuracy and strong correlation between the final SCD and the observed data. Additionally, the average value of final SCD variability (α) stands at 1.064, indicating a slight overestimation, possibly influenced by the high station density within the study area. Conversely, the mean value-based β is exceptionally low at 0.5×10^{-6}, explaining the precise estimation of precipitation amounts, as depicted in Figure 11.

4.4.2 | Evaluation of the final SCD's temporal accuracy

It is essential to assess and evaluate the stability of the final SCD's quality over the analysis period (Figure 12). The KGE″ and the CC do not exhibit any significant temporal

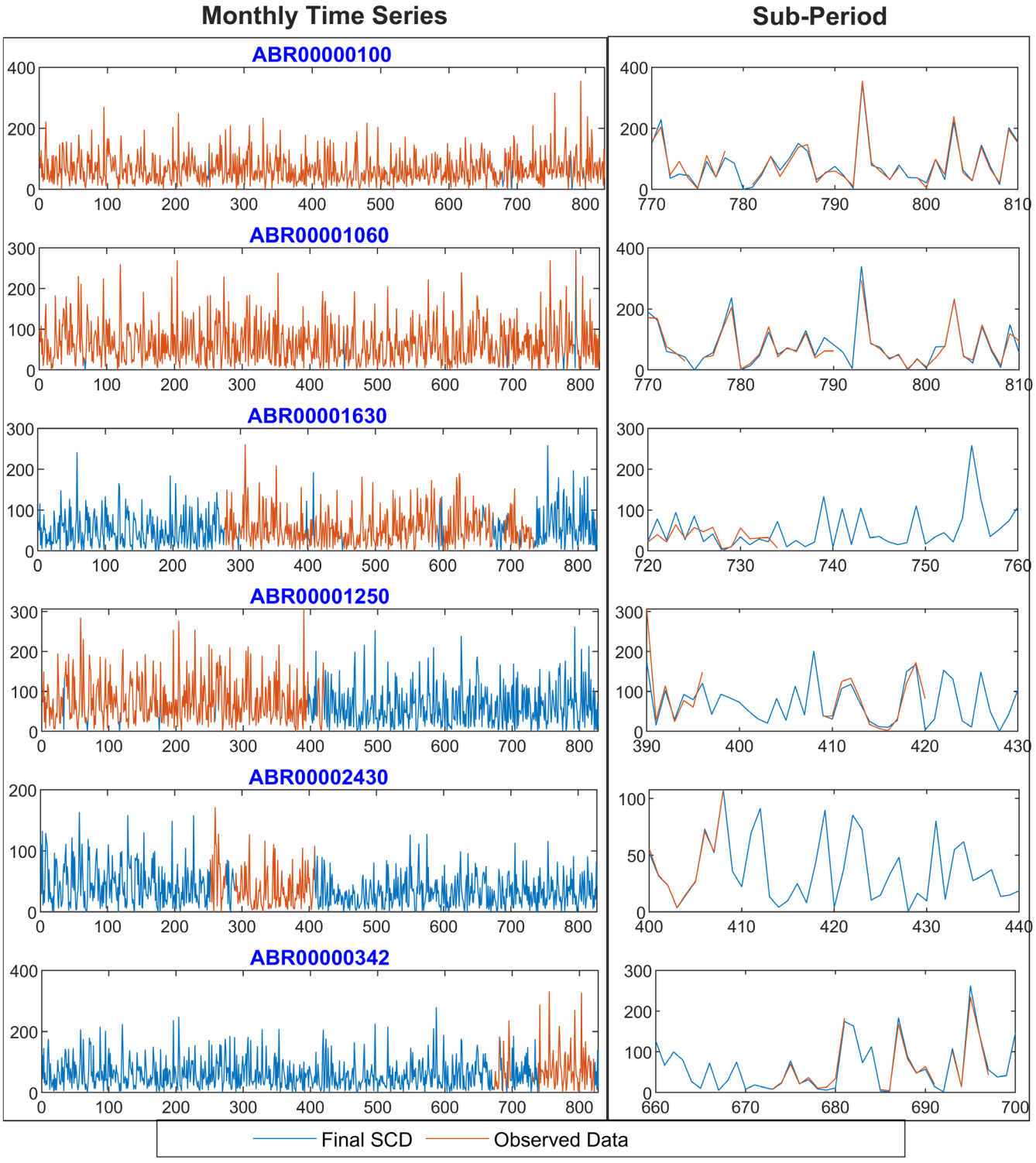

FIGURE 10 Left panel: Monthly observed and estimated precipitation series for six stations in different locations with different MV% and various gap lengths. Right panel: A subperiod of 40 months.

fluctuations from the beginning of the study period to 1973, where the yearly median KGE″ remains relatively consistent, ranging from 0.89 to 0.93 and from 0.95 to 0.98 for CC. Concurrently, the MV% follows a similar trend by representing low fluctuation and ranges between 34.34% and 40.48% over the same timeframe.

In contrast, from 1974 to the end of the study period, the yearly median KGE″ and the CC show a decreasing trend, ranging from 0.74 to 0.91 and 0.83 to 0.97, respectively. Nonetheless, the yearly MV% for the same period exhibits an opposing upward trajectory, climbing from 12.4% to 93.75%. This suggests a clear inverse

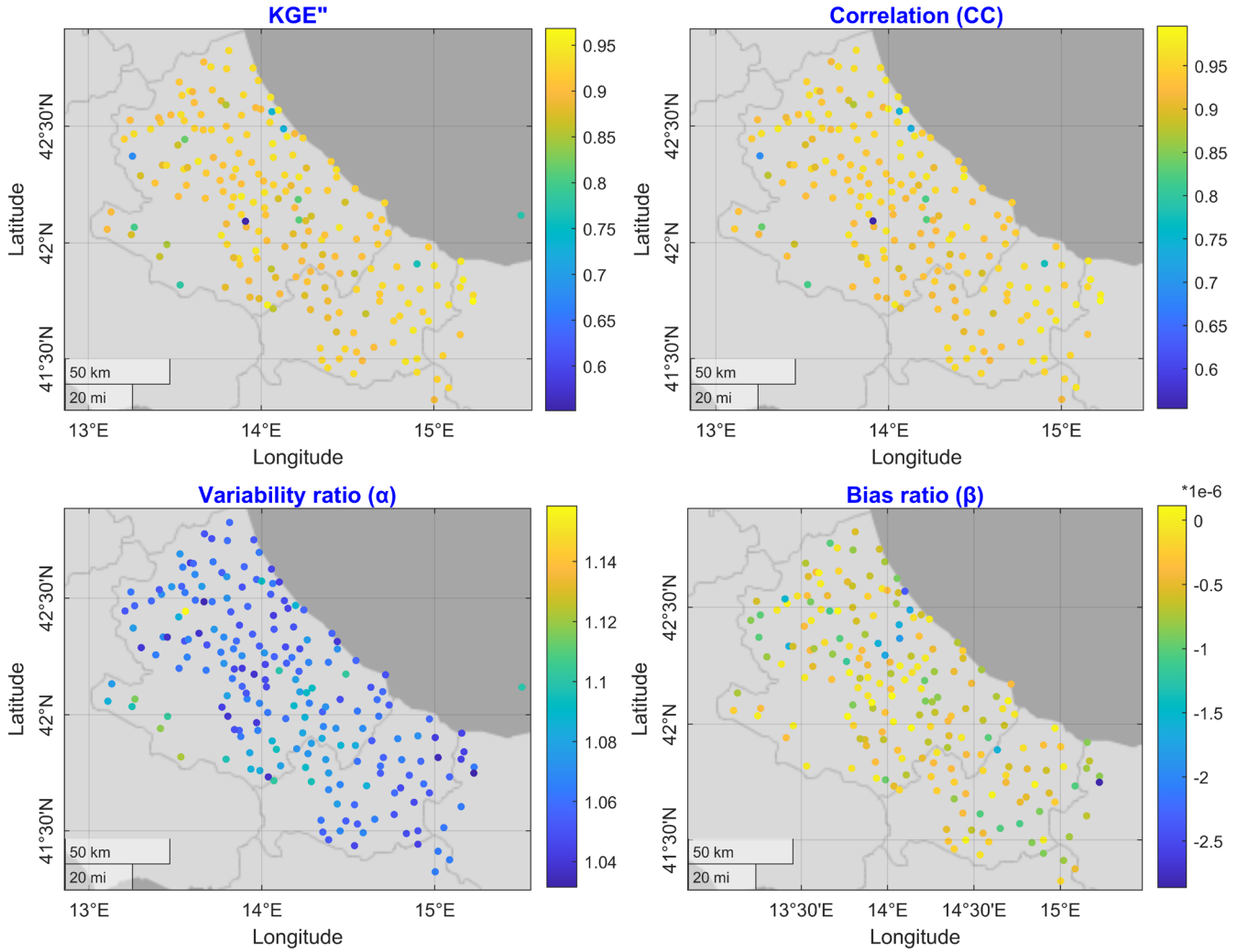

FIGURE 11 The spatial distributions of KGE″ and its three elements for the final SCD. KGE″, CC, and α all have optimal values of 1, whereas β has an optimal value of zero.

relationship between the final output's accuracy and the MV%.

The values of α and β generally hover around 1.064 and 0.0003, respectively, indicating a slight inclination of the final SCD to overestimate variability and precipitation amounts. This provides evidence that KGE is strongly influenced by CC. Hence, the year 2019 registers the lowest values for both α and β due to its highest MV%.

5 | STUDY LIMITATIONS

The recorded dataset and the reanalysed ERA5 data are essential prerequisites for the proposed gap-filling process. However, various non-climatic factors, such as station relocations, instrument types and exposure changes, land use and land cover alterations, and historical events, can potentially affect the recorded precipitation data. Moreover, unfortunately, the reanalysed ERA5 data demonstrates a limited correlation with the observed data due to the area's intricate topography and limited pixel coverage. One notable limitation arises from the reconstruction's dependence on original data, particularly during periods with numerous missing values. For example, between 2015 and 2019, many stations may have been closed due to resource or labour shortages. Despite these challenges, the methodology employed in this study effectively reconstructed missing values in the daily precipitation series, offering reliability while preserving all the statistical characteristics of the time series.

6 | CONCLUSION

In conclusion, this research outlined a methodology for constructing a densely populated spatial database of daily precipitation datasets in Central Italy, encompassing 205 stations spanning the period from 1951 to 2020, characterized by varying levels of missing values and gap durations.

Rigorous quality control measures were applied to the raw data, resulting in the successful qualification of 203 stations, with an impressive 73% of these stations retaining all their data without quality-related losses.

The proposed gap-filling strategies relied on neighbouring time series to generate estimations, which were subsequently combined to form the estimated serially complete dataset (SCD). Evaluation using various statistical metrics such as CC, RMSE, Bias%, and KGE revealed that the LSTM-based strategy outperformed others, while the QMR strategy exhibited subpar individual performance. Additionally, the estimated SCD demonstrated a strong correlation with the observed data.

The final SCD was derived by substituting estimations in the estimated SCD with observed data wherever available, resulting in a total of 201 stations. Notably, the medians of CC, KGE, α, and β for the final SCD were found to be 0.9, 0.93, 1.064, and 4.98×10^{-7}, respectively, underscoring the exceptional accuracy of the final SCD both in terms of spatial and temporal fidelity.

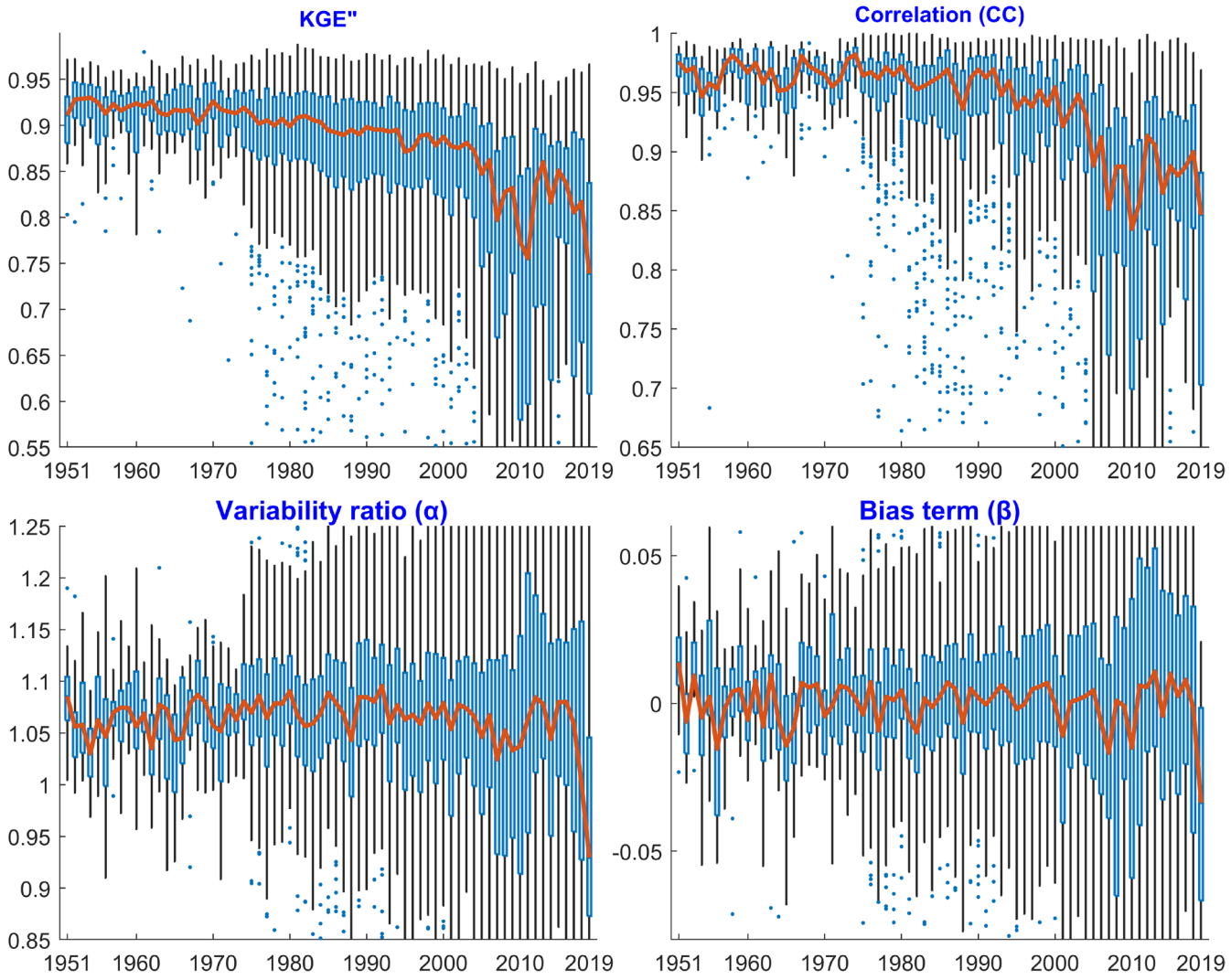

FIGURE 12 The temporal distribution of KGE″ and its three components (CC, α, and β). The four variables are computed for all stations for every year of the study period and then plotted in a box plot format (blue) with the year's median value (red line).

Study	Covered region	Covered period	No. of stations
This study	Central Italy	1951–2019	201
Curci et al. (2021), pp. 1930–2019	Central Italy	1930–2019	85
Gentilucci et al. (2018)	Central Italy	1931–2014	118
Simolo et al. (2010)	Northern Italy	1916–2004	36

TABLE 2 Comparative overview of area of interest, coverage duration, and outcome stations across various studies, including local research within the same or different regions of Italy.

This research contributes to enhancing the density of daily datasets in the Mediterranean region, a recognized hotspot for climate change. It achieves this by utilizing a significantly larger number of stations compared to other studies (Table 2), such as the SC-Earth dataset, which only incorporates some public stations within the same proposed study area. Furthermore, this study demonstrates the powerful performance of the SC-Earth methodology for gap filling in complex regions characterized by diverse topography and distinct climatic zones, including mountainous and coastal areas.

ACKNOWLEDGEMENTS

Authors acknowledge with gratefulness Regione Abruzzo-Servizio Idrografico for supplying the data analysed in this research.

FUNDING INFORMATION

This work was supported by Programma Nazionale FSE-FESR Ricerca e Innovzione 2014-2020 (PON) PhD fellowship: DOT1753918.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at https://zenodo.org/doi/10.5281/zenodo.12180685.

DATA AVAILABILITY STATEMENT

The dataset is available at https://zenodo.org/records/12180686 and comprises four files. The CSV file contains essential station information, such as station ID, latitude, longitude, elevation, and station name. The observed, estimated, and final SCD are available in MAT format, with each column in these files representing a time series of a station in the same order as in the CSV file.

ORCID

Gamal AbdElNasser Allam Abouzied https://orcid.org/0009-0000-9360-8725

REFERENCES

Aguilera, H., Guardiola-Albert, C. & Serrano-Hidalgo, C. (2020) Estimating extremely large amounts of missing precipitation data. *Journal of Hydroinformatics*, 22(3), 578–592.

Alexander, L.V., Zhang, X., Peterson, T.C., Caesar, J., Gleason, B., Klein Tank, A. et al. (2006) Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, 111, D05109.

Beck, H.E., van Dijk, A.I., Levizzani, V., Schellekens, J., Miralles, D.G., Martens, B. et al. (2017) MSWEP: 3-hourly 0.25 global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, 21(1), 589–615.

Beck, H.E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A.I., Weedon, G.P. et al. (2017) Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, 21(12), 6201–6217.

Beck, H.E., Wood, E.F., Pan, M., Fisher, C.K., Miralles, D.G., van Dijk, A.I. et al. (2019) MSWEP V2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, 100(3), 473–500.

Bellido-Jiménez, J.A., Gualda, J.E. & García-Marín, A.P. (2021) Assessing machine learning models for gap filling daily rainfall series in a semiarid region of spain. *Atmosphere.*, 12(9), 1158.

Cannon, A.J., Sobie, S.R. & Murdock, T.Q. (2015) Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? *Journal of Climate*, 28(17), 6938–6959.

Chinasho, A., Bedadi, B., Lemma, T., Tana, T., Hordofa, T. & Elias, B. (2021) Evaluation of seven gap-filling techniques for Daily Station-based rainfall datasets in South Ethiopia. *Advances in Meteorology*, 2021, 15.

Cressman, G.P. (1959) An operational objective analysis system. *Monthly Weather Review*, 87(10), 367–374.

Curci, G., Guijarro, J.A., Di Antonio, L., Di Bacco, M., Di Lena, B. & Scorzini, A.R. (2021) Building a local climate reference dataset: application to the Abruzzo region (Central Italy), 1930–2019. *International Journal of Climatology*, 41(8), 4414–4436.

Daly, C., Neilson, R.P. & Phillips, D.L. (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology and Climatology*, 33(2), 140–158.

Di Piazza, A., Conti, F.L., Noto, L.V., Viola, F. & La Loggia, G. (2011) Comparative analysis of different techniques for spatial interpolation ofrainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. *International Journal of Applied Earth Observation and Geoinformation*, 13(3), 396–408.

Durre, I., Menne, M.J., Gleason, B.E., Houston, T.G. & Vose, R.S. (2010) Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, 49(8), 1615–1633.

ECMWF. (2017) *Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate*. Copernicus Climate Change Service Climate Data Store (CDS), date of access.

Eischeid, J.K., Pasteris, P.A., Diaz, H.F., Plantico, M.S. & Lott, N.J. (2000) Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *Journal of Applied Meteorology and Climatology*, 39(9), 1580–1591.

Garcia-Marin, A., Jiménez-Hornero, F.J. & Ayuso-Munoz, J. (2008) Universal multifractal description of an hourly rainfall time series from a location in southern Spain. *Atmosfera*, 21(4), 347–355.

Gentilucci, M., Barbieri, M., Burt, P. & D'Aprile, F. (2018) Preliminary data validation and reconstruction of temperature and precipitation in Central Italy. *Geosciences*, 8(6), 202.

Grillakis, M.G., Polykretis, C., Manoudakis, S., Seiradakis, K.D. & Alexakis, D.D. (2020) A quantile mapping method to fill in discontinued daily precipitation time series. *Watermark*, 12(8), 2304.

Gupta, H.V., Kling, H., Yilmaz, K.K. & Martinez, G.F. (2009) Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91.

Hamada, A., Arakawa, O. & Yatagai, A. (2011) An automated quality control method for daily rain-gauge data. *Global Environmental Research*, 15(2), 183–192.

Heo, J.-H., Ahn, H., Shin, J.-Y., Kjeldsen, T.R. & Jeong, C. (2019) Probability distributions for a quantile mapping technique for a bias correction of precipitation data: a case study to precipitation data under climate change. *Watermark*, 11(7), 1475.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049.

Ilunga, M. & Stephenson, D. (2005) Infilling streamflow data using feed-forward back-propagation (BP) artificial neural networks: application of standard BP and pseudo mac laurin power series BP techniques. *Water SA*, 31(2), 171–176.

Kemp, W., Burnell, D., Everson, D. & Thomson, A. (1983) Estimating missing daily maximum and minimum temperatures. *Journal of Applied Meteorology and Climatology*, 22(9), 1587–1593.

Kling, H., Fuchs, M. & Paulin, M. (2012) Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424, 264–277.

Larson, L.W. & Peck, E.L. (1974) Accuracy of precipitation measurements for hydrologic modeling. *Water Resources Research*, 10(4), 857–863.

Leopardi, M. & Scorzini, A.R. (2015) Effects of wildfires on peak discharges in watersheds. *iForest - Biogeosciences and Forestry*, 8(3), 302–307.

Lloyd, C. (2005) Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain. *Journal of Hydrology*, 308(1–4), 128–150.

Lo Presti, R., Barca, E. & Passarella, G. (2010) A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental Monitoring and Assessment*, 160(1), 1–22.

Longman, R.J., Frazier, A.G., Newman, A.J., Giambelluca, T.W., Schanzenbach, D., Kagawa-Viviani, A. et al. (2019) High-resolution gridded daily rainfall and temperature for the Hawaiian islands (1990–2014). *Journal of Hydrometeorology*, 20(3), 489–508.

Mwale, F.D., Adeloye, A.J. & Rustum, R. (2012) Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach. *Physics and Chemistry of the Earth, Parts A/B/C.*, 50–52, 34–43.

Newman, A.J., Clark, M.P., Craig, J., Nijssen, B., Wood, A., Gutmann, E. et al. (2015) Gridded ensemble precipitation and temperature estimates for the contiguous United States. *Journal of Hydrometeorology*, 16(6), 2481–2500.

Nkuna, T. & Odiyo, J.O. (2011) Filling of missing rainfall data in Luvuvhu RiverCatchment using artificial neural networks. *Physics and Chemistry of the Earth, Parts A/B/C*, 36(14–15), 830–835.

Nonhebel, S. (1993) *The importance of weather data in crop growth simulation models and assessment of climatic change effects*. Wageningen: Wageningen University and Research.

Prudhomme, C., Dadson, S., Morris, D., Williamson, J., Goodsell, G., Crooks, S. et al. (2012) Future flows climate: an ensemble of 1-km climate change projections for hydrological application in Great Britain. *Earth System Science Data*, 4(1), 143–148.

Qin, Y., Ren, G., Zhang, P., Wu, L. & Wen, K. (2021) An imputation method for the climatic data with strong seasonality and spatial correlation. *Theoretical and Applied Climatology*, 144(1), 203–213.

Quagraine, K.A., Nkrumah, F., Klein, C., Klutse, N.A.B. & Quagraine, K.T. (2020) West African summer monsoon precipitation variability as represented by reanalysis datasets. *Climate*, 8(10), 111.

Santos, L., Thirel, G. & Perrin, C. (2018) Pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, 22(8), 4583–4591.

Shepard, D. (1968) *A two-dimensional interpolation function for irregularly-spaced data*, (pp. 517–524). Presented at the Proceedings of the 1968 23rd ACM national conference.

Simolo, C., Brunetti, M., Maugeri, M. & Nanni, T. (2010) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *International Journal of Climatology*, 30(10), 1564–1576.

Tang, G., Clark, M.P., Newman, A.J., Wood, A.W., Papalexiou, S.M., Vionnet, V. et al. (2020) SCDNA: a serially complete precipitation and temperature dataset for North America from 1979 to 2018. *Earth System Science Data*, 12(4), 2381–2409.

Tang, G., Clark, M.P. & Papalexiou, S.M. (2021) SC-earth: a station-based serially complete earth dataset from 1950 to 2019. *Journal of Climate*, 34(16), 6493–6511.

Tarek, M., Brissette, F.P. & Arsenault, R. (2020) Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. *Hydrology and Earth System Sciences*, 24(5), 2527–2544.

Teegavarapu, R.S. & Chandramouli, V. (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, 312(1–4), 191–206.

Vieux, B.E. (2001) *Distributed hydrologic modeling using GIS*. Cherokee: Springer, pp. 1–17.

How to cite this article: Abouzied, G.A.A., Tang, G., Papalexiou, S.M., Clark, M.P., Aruffo, E. & Di Carlo, P. (2025) Completion of the Central Italy daily precipitation instrumental data series from 1951 to 2019. *Geoscience Data Journal*, 12, e267. Available from: https://doi.org/10.1002/gdj3.267