Missing Data Imputation for Multisite Rainfall Networks: A Comparison between Geostatistical Interpolation and Pattern-Based Estimation on Different Terrain Types

FABIO ORIANI

Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, University of Lausanne, Lausanne, Switzerland

SIMON STISEN

Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark

MEHMET C. DEMIREL

Department of Civil Engineering, Istanbul Technical University, Istanbul, Turkey

GREGOIRE MARIETHOZ

Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, University of Lausanne, Lausanne, Switzerland

(Manuscript received 19 September 2019, in final form 7 July 2020)

ABSTRACT

Missing rainfall data are a major limitation for distributed hydrological modeling and climate studies. Practitioners need reliable approaches that can be employed on a daily basis, often with too limited data in space to feed complex predictive models. In this study we compare different automatic approaches for missing data imputation, including geostatistical interpolation and pattern-based estimation algorithms. We introduce two pattern-based approaches based on the analysis of historical data patterns: (i) an iterative version of *K*-nearest neighbor (IKNN) and (ii) a new algorithm called vector sampling (VS) that combines concepts of multiple-point statistics and resampling. Both algorithms can draw estimations from variably incomplete data patterns, allowing the target dataset to be at the same time the training dataset. Tested on five case studies from Denmark, Australia, and Switzerland, the algorithms show a different performance that seems to be related to the terrain type: on flat terrains with spatially homogeneous rain events, geostatistical interpolation tends to minimize the average error, while in mountainous regions with nonstationary rainfall statistics, data mining can recover better the rainfall patterns. The VS algorithm, requiring minimal parameterization, turns out to be a convenient option for routine application on complex and poorly gauged terrains.

KEYWORDS: Hydrometeorology; Numerical analysis/modeling; Pattern detection; Statistical techniques; Hydrologic models

1. Introduction

A realistic spatiotemporal representation of rainfall is of primary importance to estimate the uncertainty of basin recharge and its propagation to the surface and underground circulation. Hydrologists can benefit from different sources of information, such as remote sensing data products, but ground-based rain gauge measurements remain the most common and reliable data source. A number of comparative studies considering remote sensing products and rain gauge networks (Guilloteau et al. 2016; He et al. 2018; Seo et al. 2018; Bayabil et al. 2019; Lasser et al. 2019) demonstrate the critical role of a sufficiently dense ground measurement network.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-19-0220.s1.

Corresponding author: Fabio Oriani, fabio.oriani@protonmail.com

DOI: 10.1175/JHM-D-19-0220.1

This information is important to detect the rainfall intensity pattern and develop efficient bias-correction techniques of radar and satellite estimations. A primary hurdle is the presence of missing data, which especially affects large historical datasets. Typical causes are measurement failure, temporary inactivity of stations, and changes in the configuration of the network. However, of even greater concern is the general global trend of diminishing rain gauge networks. Following a steady increase in global rain gauge measurements up until the 1970s there has been a rapid decline, exemplified by the Global Precipitation Climatology Project/Global Precipitation Climatology Centre (GPCP/GPCC) database which has declined from 38 000 to 7000 gauges from 1988 to 1999 (New et al. 2001). These issues do not only limit the information content of the datasets and their use, but can also generate inconsistencies in subsequent modeling steps, such as calibration of hydrological models or merging with other data products. Therefore, infilling missing data gaps (also called reconstruction, imputation, or patching) has become a routine preprocessing phase of hydrological studies.

A number of geostatistical techniques have been developed through the last decades to estimate missing rainfall data. The main principle of these approaches is the statistical correlation among precipitation data at different locations, which reflects the physical spatiotemporal continuity of rainfall. This allows using nearby measurements in space and time to predict missing data with uncertainty. Standard gap-filling procedures (ASCE 1996) are based on the linear combination of measures at correlated (predictor) stations to estimate the missing datum at one target station or ungauged location, within the same or nearby time steps. Different types of linear combination coefficients have been proposed, the most common ones being: the ratio of the expected value between each predictor and the target, computed on the historical record (Linsley et al. 1988; Azman et al. 2015), a function of the station distance (inverse distance weighting), and the linear correlation estimated on their time series (Wei and McGuinness 1973; Cooke and Mostaghimi 1992; Mair and Fares 2010; Noori et al. 2014; Caldera et al. 2016; Moeletsi et al. 2016; Ismail and Ibrahim 2017; Woldesenbet et al. 2017). Other types of combination include the geometric median (Burhanuddin et al. 2015).

Geostatistical approaches have allowed a more explicit mathematical formulation of the spatiotemporal behavior of rainfall: the most common one is kriging, which models the observed spatial variability within the dataset by means of the variogram function, then used to estimate the weights for each predictor station. Kriging has been applied to the estimation of missing rainfall data, with the possibility of incorporating auxiliary variables to improve the predictions (Di Piazza et al. 2011; Verworn and Haberlandt 2011; Kisaka et al. 2016). The advantages of this approach in comparison to its simpler predecessors are the estimation of uncertainty using the local kriging variance, its minimization, and the use of a specific spatial model, possibly including auxiliary variables. Nevertheless, the technique relies on some strong statistical hypothesis (e.g., Gaussianity) that require histogram transformation on the target variable to be satisfied (Benoit et al. 2018a). More importantly, the quantification of the spatial structure is only based on low-order moments (local mean and variance) that may not represent the complexity of rainfall. A step forward in this sense has been done with the proposition of multivariate copula functions (Bárdossy and Pegram 2014), allowing a more precise estimation of the rainfall probability distribution in space. Further modeling efforts are needed to preserve high-order statistical features of rainfall, e.g., anisotropy, connectivity structure, the presence of dry regions, and the variability of rainfall types (Benoit et al. 2018b). Proposed modeling approaches, such as machine learning (Kim and Pachepsky 2010; Kajornrit et al. 2012a), fuzzy logic (Abebe et al. 2000; Kajornrit et al. 2012b), or nonlinear mathematical programming (Teegavarapu 2012), based on calibration, allow introducing more flexibility in the statistical dependencies and preserving complex behaviors, but they generally require case-dependent setup and a careful choice of the model structure.

Given the complexity of a full geostatistical analysis, practitioners tend to rely on inverse distance weighting (ID) interpolation for routine gap filling, the main reasons being its simplicity and rapidity of use, and robustness in automatic applications, especially in sparse-gaps scenarios, with results often comparable to more complex estimations (Tung 2013; Yang et al. 2015; Caldera et al. 2016). These techniques often ignore the full time–space dependence structure, simplifying the model according to the scale investigated. Strong limitations of ID are the tendency to smooth out the spatial variations, the inability to generate values outside the range of their neighbors in space, and the ignorance of complex spatial features observed in data patterns. Attention for these features emerged in the early 2000s, in particular regarding methods for generating space–time rainfall data that preserve complex features (Grayson et al. 2002). In Teegavarapu and Chandramouli (2005), an improvement of inverse distance weighting is obtained based on temporal data-pattern analysis: the predictor stations are chosen based on the similarity in the discretized local rainfall trend. Another proposed strategy (Hema and Kant 2017) aims at reconstructing hourly missing data by categorizing rainfall into event types and using a sliding-window approach to detect time periods where events occur. This approach presents the

relevant property to preserve more complex data aggregation and variability but involves a complex set of rules to define the sliding window size in space and time. Moreover, it is only applicable on a time scale where continuity of rainfall is expected. As an alternative strategy, some authors proposed to incorporate gridded rainfall products derived from satellite data into rain gauge interpolation, obtaining a better preservation of the spatial distribution at the regional scale (Huffman et al. 1995; Wagner et al. 2012; Stisen and Tumbo 2015). Spatial rainfall patterns have been related to the hydrological response: in the analysis of simulated land surface variables, rainfall has been identified as a primary influencing factor for the spatial patterns of land surface temperature and evapotranspiration (Koch et al. 2017). Different spatial rainfall products have been analyzed in relation to the spatial hydrological response (He et al. 2018) and used to calibrate hydrological models (Demirel et al. 2018). These works stress the importance of data-pattern analysis to estimate a realistic spatial distribution of rainfall.

In this paper we compare geostatistical and pattern-based estimation approaches for missing-data imputation. Among pattern-based techniques, we propose two strategies based on spatial data patterns, which consider more high-order features than geostatistical interpolation techniques, but at a comparable cost in terms of complexity and computation. The first technique proposed, called vector sampling (VS), is inspired by algorithms belonging to the geostatistical family of multiple-point statistics (MPS) (Strebelle 2002), developed to simulate complex heterogeneity based on data-pattern generation. MPS techniques use a training image (TI) that constitutes a catalog of data patterns representing the variable to simulate, providing information about high-order statistical dependencies and complex features such as geometry, connectivity, and orientation. In the case of rainfall data, a historical multisite dataset can be used as TI. Among MPS techniques, the direct sampling algorithm (DS) (Mariethoz et al. 2010), based on sampling random data from the TI, has been already applied to rainfall simulation (Oriani et al. 2014, 2017b) and streamflow gap filling (Oriani et al. 2016; Dembélé et al. 2019). With a relatively simple parameterization, DS preserves complex rainfall features by generating similar data patterns as the ones found in the TI. Nevertheless, at the present stage of development, it cannot handle efficiently the presence of numerous data gaps inside the TI, since it requires additional parameterization to penalize incomplete patterns. Moreover, the algorithm aims at studying the uncertainty at multiple scales by considering patterns of different size and allowing unconditional simulation. This comes at a cost in terms of computational time and parameterization complexity. When dealing with missing data in multisite measurement networks, this complexity is usually unnecessary since regional and long-term fluctuations are already determined by the available data from other stations.

Another pattern-based technique considered in this paper is *K*-nearest neighbor (KNN) estimation. This type of strategy, already applied to multisite rainfall estimation (Apipattanavis et al. 2007; Caraway et al. 2014) and forecast (Wu 2009; Hu et al. 2013), relies on the use of nearby-station measurements, aggregated statistics, or other predictor variables to identify similar rainfall patterns in the historical record. The associated rainfall amounts at the station of interest are then randomly sampled or used for conditional inference. These techniques are not usually suitable to handle variable missing-data configurations, since they train the model with a fixed set of predictor/target variables. We propose here a combination of KNN with an iterative sampling strategy (IKNN) to allow a variable missing-data configuration for every day, and consider it as an alternative to VS.

A total of seven algorithms, including ID, four variants of kriging, and the resampling techniques VS and IKNN, are compared by performing a series of cross-validation exercises. Five study cases are considered, from Denmark, Switzerland, and Australia, presenting different terrain complexity and climate settings. Sensitivity to the amount of missing data is also analyzed by generating different missing-data scenarios. These scenarios mimic the closure of stations for multiple days to months. Unlike previously proposed studies, missing-data percentages ranging from 20% to 80% are considered, which represent cases where large parts of the network are cut out for continuous periods of time. Importantly, the training datasets naturally contain missing data, which brings the study really close to real-case conditions. This is particularly interesting to explore the performance of the proposed pattern-based techniques in handling incomplete predictor patterns, a situation which is usually not optimally handled by this kind of techniques.

The paper is organized as follows: the compared techniques are described in sections 2 and 3 contains information about the datasets used, section 4 shows the experiment design, and the results are presented in section 5. A discussion on the results and final remarks are given in sections 6 and 7, respectively.

2. Methodology

The techniques presented in this study are applied to the estimation of missing daily rainfall data in a multisite network. As suggested by the weak autocorrelation observed in daily time series (Oriani et al. 2017a), at this time scale, spatial relations among stations are more

relevant to predict missing data than temporal relations among data from the same station. Therefore, the compared techniques use spatial data as predictor variables to estimate missing data in the same day.

a. *Inverse distance weighting interpolation*

ID interpolation estimates the daily rainfall r at any location x_0 as the weighted linear combination:

$$\hat{r}(x_0) = \left(\sum_{s=1}^{S} d_s^{-\Theta} \right)^{-1} \sum_{s=1}^{S} d_s^{-\Theta} r_s, \tag{1}$$

where r_s is the rainfall amount observed at the sth station in the current day, with $s = 1, \ldots, S$ being the set of informed locations in the current day. Parameter d_i is the Euclidean distances between the ith and the target locations. For the extension of the regions considered in this study, we can avoid defining a limited neighborhood for the predictor data. The only parameter in the ID formulation used here is the power Θ applied to the distance. After preliminary tests (attached as supplemental material), $\Theta = 5$ has been chosen for all study cases.

b. *Ordinary kriging and indicator kriging*

Ordinary kriging (OK), introduced in the early 1960s by George Matheron (Matheron 1963), constitutes the basis for geostatistical interpolation with uncertainty estimation. It makes use of a parametric covariance model, formalized in the variogram function, to make predictions as a function of data. Following the notation of section 2a, the estimation of r at any locations x_0 is formulated as the weighted mean:

$$\hat{r}(x_0) = \sum_{s=1}^{S} w_s r_s \tag{2}$$

where w_s is the weight associated to the datum r_s. The weights $\mathbf{w}$ are computed by solving the kriging system of the form:

$$\begin{bmatrix} \mathbf{w} \\ \lambda \end{bmatrix} = \begin{bmatrix} \Gamma(x_s, x_{s'}) & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \boldsymbol{\gamma}(x_s, x_0) \\ 1 \end{bmatrix}, \tag{3}$$

where λ is a Lagrangian multiplier, $\boldsymbol{\Gamma}(x_s, x_{s'})$ is the modeled variogram matrix for all given S data with coordinates $x_s, x_{s'}$, with both s and s' taking values in $1, \ldots, S$. The vector $\mathbf{1}$ is a ones vector of size N, and $\boldsymbol{\gamma}(x_s, x_0)$ is a vector of modeled variogram given data at space coordinates x_s and the target estimation point coordinate x_0. Spatial variability is controlled by the chosen variogram model $\boldsymbol{\gamma}(\cdot, \cdot)$. Following (Cressie 1985), two common types of isotropic variogram models are used in this study. The first one, chosen for its simplicity and robustness, is the linear model with no nugget effect $\boldsymbol{\gamma}(x_s, x_{s'}) = a|(x_s - x_{s'})|$, that imposes a constant variation of the modeled quantity as a function of distance, with the sole parameter a. The estimation weights $\mathbf{w}$ and ultimately the kriging mean [Eq. (2)] do not depend on a, but only the kriging variance does. Since in this study only the kriging mean is used, a can be given an arbitrary real value in (0, inf). The second variogram model considered, capable of modeling sharper variations, is the exponential one $\boldsymbol{\gamma}(x_s - x_{s'}) = a_1 + a_2[1 - \exp(-|x_s - x_{s'}|/a_3)]$, where $\mathbf{a} = a_1, a_2, a_3$; is the parameter vector. The variogram model is fitted to the sample variogram of the day (calibration of $\mathbf{a}$) with a L1 norm minimization scheme. OK provides unbiased estimation and minimization of the error variance but relies on the hypothesis of stationarity and requires a symmetrical probability distribution for the modeled variable. These conditions are typically not satisfied for daily rainfall, exhibiting a dry–wet pattern and asymmetric distribution. Nevertheless, in the case of missing data, OK is applied as is if the amount of available data in space does not allow building a more complex model.

In this study, we also consider the case where OK is applied in a more complex framework, accommodating nonstationarity and an asymmetric probability distribution. First, a quantile–quantile transformation is applied to the nonzero rainfall amount to obtain a normal marginal distribution. Then, as done in Chappell et al. (2013), we estimate missing rainfall data in two steps: first we apply OK to the indicator variable for the dry/wet pattern. In this technique, called indicator kriging (IK), the indicator variable $b = \mathbb{I}_{r>0} \in \{0, 1\}$ (0 = dry, 1 = wet) constitutes the available occurrence data in space for the day. Similar to Eq. (2), OK is applied to estimate $\hat{b}_{\mathrm{OK}}$ at unknown locations, using the linear variogram model. As a result $\hat{b}_{\mathrm{OK}}$ takes real values in [0, 1], that can be interpreted as the estimated occurrence probability at each unknown locations. Estimated wet occurrence is obtained as $\hat{b} = \mathbb{I}_{\hat{b}_{\mathrm{OK}} > \overline{b}}$, i.e., if $\hat{b}_{\mathrm{OK}}$ exceeds the average occurrence $\overline{b}$ observed in space. When the dry/wet pattern estimation is obtained, the rainfall amount at wet locations ($\hat{b} = 1$) is estimated using OK. Wet-locations amount is then transformed back from normal to the initial rainfall distribution. The kriging interpolations are performed with the python package PyKrige (https://pypi.org/project/PyKrige/).

c. *Iterative K-nearest neighbor*

KNN is a strategy for nonparametric inference (Stone 1977) that sees application in regression, classification, and imputation problems. Conversely to ID and OK, which can estimate a datum anywhere in space as a weighted mean from data at neighbor locations, the KNN algorithm used here applies a weighted mean

from a sample of historical data at the target location. This requires the presence of historical data at the multisite station network considered. Let us define a dataset composed by a variable r defined for $\mathbf{s} = 1, \ldots, S$ stations and $\mathbf{i} = 1, \ldots, I$ days. For a target station $(i0, s0)$ outside (i, s), let us consider a predictor or conditioning pattern $\mathbf{r}_{i0,s}$, composed by data from all s stations in the day $i0$. Each daily pattern $\mathbf{r}_{i,s}$ is compared with $\mathbf{r}_{i0,s}$ by computing a distance measure $d_i|\mathbf{r}_{i0,s}$, i.e., a measure of dissimilarity for the ith day given $\mathbf{r}_{i0,s}$. For this purpose, the commonly used Euclidean norm is adopted:

$$d_i|\mathbf{r}_{i0,s} = \left[\sum_{s=1}^{S}\left(r_{i,s} - r_{i0,s}\right)^2\right]^{1/2}. \tag{4}$$

Historical data at the target station $\mathbf{r}_{s0}$ are then ranked by increasing d:

$$\mathbf{r}_{s0}|\mathbf{d} \rightarrow \mathbf{r}_k = \{r_1, \ldots, r_k, \ldots, r_I : d_k \leq d_{k+1}\}. \tag{5}$$

The target datum is then estimated as the weighted mean:

$$\hat{r}_{i0,s0} = \left(\sum_{k=1}^{K} d_k^{-1}\right)^{-1} \sum_{k=1}^{K} d_k^{-1} r_k, \tag{6}$$

where K is the parameter indicating the number of nearest neighbors considered for the estimation, i.e., the K historical data at s_0 that present the most similar data pattern at other stations.

When missing data exist in $r_{i,s}$ as well, computing the pattern distance as in Eq. (4) is not possible. The original KNN algorithm is therefore not applicable to estimate sparse missing data into a multisite network. To overcome this limitation, we propose an IKNN algorithm, where the KNN estimator is inserted into an iterative workflow for multivariate imputation (Rubin 2004; van Buuren and Groothuis-Oudshoorn 2010). Let us consider the situation where the target data to estimate are any missing data inside (s, i). At the beginning, they are filled with the arithmetic mean of all historical data from the same station s:

$$\hat{r}_s = \left(\sum_{i\in\mathbf{i}} \mathbb{I}_{r_{i,s}\neq \mathrm{NaN}}\right)^{-1} \sum_{\substack{i\in\mathbf{i},\\ r_{i,s}\neq \mathrm{NaN}}} r_{i,s}, \tag{7}$$

where the indicator $\mathbb{I}_{r_{i,s}\neq \mathrm{NaN}}$ equals 1 when $r_{i,s}$ is not missing and 0 otherwise. For simplicity, let us drop i, s and define observed data over all stations and days with $\mathbf{r}_{\mathrm{obs}}$, estimated data with $\hat{r}_1$, $\hat{r}_2$, …, $\hat{r}_N$, and with knn$(\cdot)$, the KNN estimation obtained by Eqs. (4)–(6). All data initially estimated with $\hat{r}_s$ are iteratively re-estimated by applying knn$(\cdot)$ in a round-robin fashion:

$$\begin{aligned}
\hat{r}_1 &= \mathrm{knn}(1)|\mathbf{r}_{\mathrm{obs}}, \hat{r}_{2,\ldots,N},\\
\hat{r}_2 &= \mathrm{knn}(2)|\mathbf{r}_{\mathrm{obs}}, \hat{r}_{1,3,\ldots,N},\\
&\vdots\\
\hat{r}_n &= \mathrm{knn}(n)|\mathbf{r}_{\mathrm{obs}}, \hat{r}_{1,\ldots,n-1,n+1,\ldots,N}.
\end{aligned} \tag{8}$$

The whole loop is repeated m times until convergence, with suggested values of m ranging from 10 to 20 (van Buuren and Groothuis-Oudshoorn 2010). After preliminary tests, we set up $m = 10$ for all study cases. Another parameter that sensibly affects the estimation is K in Eq. (6) (see the online supplemental material). Parameter K is set to 5 for the Alice Springs study case and 10 for all other study cases (see section 3). The IKNN implementation used in this study is the python package ''sklearn.impute.IterativeImputer'' combined with the estimator ''sklearn.neighbors.KNeighborsRegressor'' (Pedregosa et al. 2011).

d. Vector sampling

We introduce here the VS algorithm for gap filling inside multisite observational networks. Similarly to KNN (section 2c), VS relies on historical data patterns with a distance-rank-estimation workflow, but it addresses the issue of incomplete data patterns differently. Following the notation of section 2c, to estimate the missing data for the day $i0$, let us define a predictor or conditioning data pattern $\mathbf{r}_{i0,s}$, composed by data from all informed s stations in the day $i0$. All days patterns present in the dataset $\mathbf{r}_{i,s}$ is compared with $\mathbf{r}_{i0,s}$ by computing the distance:

$$d_i|\mathbf{r}_{i0,s} = \sum_{\substack{s\in\mathbf{s},\\ r_{i,s}\neq \mathrm{NaN},\\ r_{i0,s}\neq \mathrm{NaN}}} (r_{i,s} - r_{i0,s})^2 + \sum_{\substack{s\in\mathbf{s},\\ r_{i,s}= \mathrm{NaN},\\ r_{i0,s}\neq \mathrm{NaN}}} \hat{d}_s|r_{i0,s}. \tag{9}$$

In this way, the candidate pattern can be composed by informed (nonmissing) stations ($r_{i,s} \neq \mathrm{NaN}$), for which the distance is similar to the one applied in KNN [Eq. (4)], and by missing-data stations ($r_{i,s} = \mathrm{NaN}$), for which the distance is estimated ($\hat{d}_s|\mathbf{r}_{i0,s}$). This estimation is given as the arithmetic mean of the squared error between $r_{i0,s}$ and the historical data from the same station s:

$$\hat{d}_s|r_{i0,s} = \left(\sum_{i\in\mathbf{i}} \mathbb{I}_{r_{i,s}\neq \mathrm{NaN}}\right)^{-1} \sum_{\substack{i\in\mathbf{i},\\ r_{i,s}\neq \mathrm{NaN}}} (r_{i0,s} - r_{i,s})^2. \tag{10}$$

Relying on the hypothesis of stationarity and independence in time, the proposed estimator considers data from the underlying marginal probability distribution of the rainfall amount associated to the s station. It is by

definition unbiased and minimizes the marginal expectation error, as observed in numerical tests (see appendix B). Note that the square root has been dropped from Eqs. (9) and (10), since, as seen in the following, the algorithm only relies on the ranking of d_i.

The computation of d is completed by imposing $d_i = +\infty$ daily patterns that presents missing data for all missing-data locations of $i0$. This operation allows discarding patterns that cannot contribute to the estimation. Then, the set of historical daily patterns is ranked by increasing d:

$$\{\mathbf{r}_{i,s}\}|\mathbf{d} \rightarrow \{\mathbf{r}_{k,s}\} = \{\mathbf{r}_{1,s}, \ldots, \mathbf{r}_{k,s}, \ldots, \mathbf{r}_{I,s} : d_k \leq d_{k+1}\}. \tag{11}$$

At this point, data at any missing stations s in the target day $i0$ are estimated as the weighted mean:

$$\hat{r}_{i0,s} = \left(\sum_{k=1}^{K} d_k^{-1} \right)^{-1} \sum_{k=1}^{K} d_k^{-1} r_{k,s}, \tag{12}$$

where K, similarly to KNN, is the parameter indicating the number of nearest neighbors considered for the estimation. As for IKNN, after preliminary tests (see supplemental material), K is set to 5 for the Alice Springs study case and 10 for all other study cases (see section 3). It can happen that all K-nearest neighbors present a missing datum for locations to be filled in $i0$. In this case, the whole workflow is repeated to fill in the remaining locations. See appendix A for the algorithm pseudocode and section 8 for the code availability information.

3. Case study data

Five case studies of daily rain gauge networks over different types of terrains are considered. Figure 1 shows the study areas with the spatial distribution of the rain gauges, and Table 1 contain a summary of the datasets and climate settings. Presented in order of terrain complexity, the study areas include:

- the Skjern catchment in western Jutland, Denmark (Fig. 1a), characterized by flat topography and mainly uniform rainfall events of stratiform type;
- the Alice Springs area in the Northern Territory, Australia (Fig. 1b), an arid region featuring erratic rainfall events;
- the Swiss Plateau, Switzerland (Fig. 1c), a flat-to-hilly region surrounded by mountain chains, that features marked warm-season thunderstorms;
- the southern outback of Tasmania, Australia (Fig. 1d), presenting a longitudinal topographic gradient that affects the intensity of rainfall;
- the Swiss region of Wallis, Switzerland (Fig. 1e), constituted of an alpine landscape that presents complex climatic and rainfall patterns strongly affected by relief and valley orientation.

The spatial extent of the study zones varies in the mesoscale range from tens to a few hundreds of kilometers in length, with a rain gauge distance that varies from one to tens of kilometers. Missing data are present in all datasets in an average amount of 20%–30%, but they are mainly concentrated in the first years for the Swiss Plateau, the Tasmanian outback, and Wallis, indicating a gradual expansion of the measurement network. Conversely, for the Skjern catchment and Alice Springs, missing data mainly affect the latest years, indicating a depletion of the network. A few cases of stations in the Skjern database, presenting more than 50% of missing data, have been removed to have more control on the presence of missing data in the cross-validation exercises (see section 4). However, all considered techniques can be applied with such scarce data records, as shown in the tests.

The data are sourced from the Danish Hydrological Observatory (http://www.hobe.dk), the Federal Office of Meteorology and Climatology Meteoswiss (https://www.meteoswiss.admin.ch/), and the Bureau of Meteorology of the Australian Government (http://www.bom.gov.au/).

4. Experiment design and validation

The experiment proposed in this study analyses the efficiency of the considered gap-filling techniques with a series of cross-validation exercises, where artificial gaps are created in the five rain gauge datasets considered. From every dataset, the 2000-day period presenting the smallest amount of missing data has been chosen for validation, using the remaining data as training dataset for the algorithms IKNN and VS. The training periods contain about 30%–40% of natural data missing, while in the validation period this amount does not surpass 17% for the Australian datasets, and it is near zero for the others (see section 3). These prior gaps are kept in both the training and validation parts of the datasets as it happens in real application.

For the cross-validation exercises, four depletion scenarios have been created, consisting of artificial missing data for approximately 20%, 40%, 60%, and 80% of all datasets. They are organized in random gaps simulating real station closure over multiple days, i.e., by placing missing data at consecutive time steps at one or multiple stations. The gap parts superposing to natural missing data are not considered in the validation. Once the gap scenarios have been created in the validation time

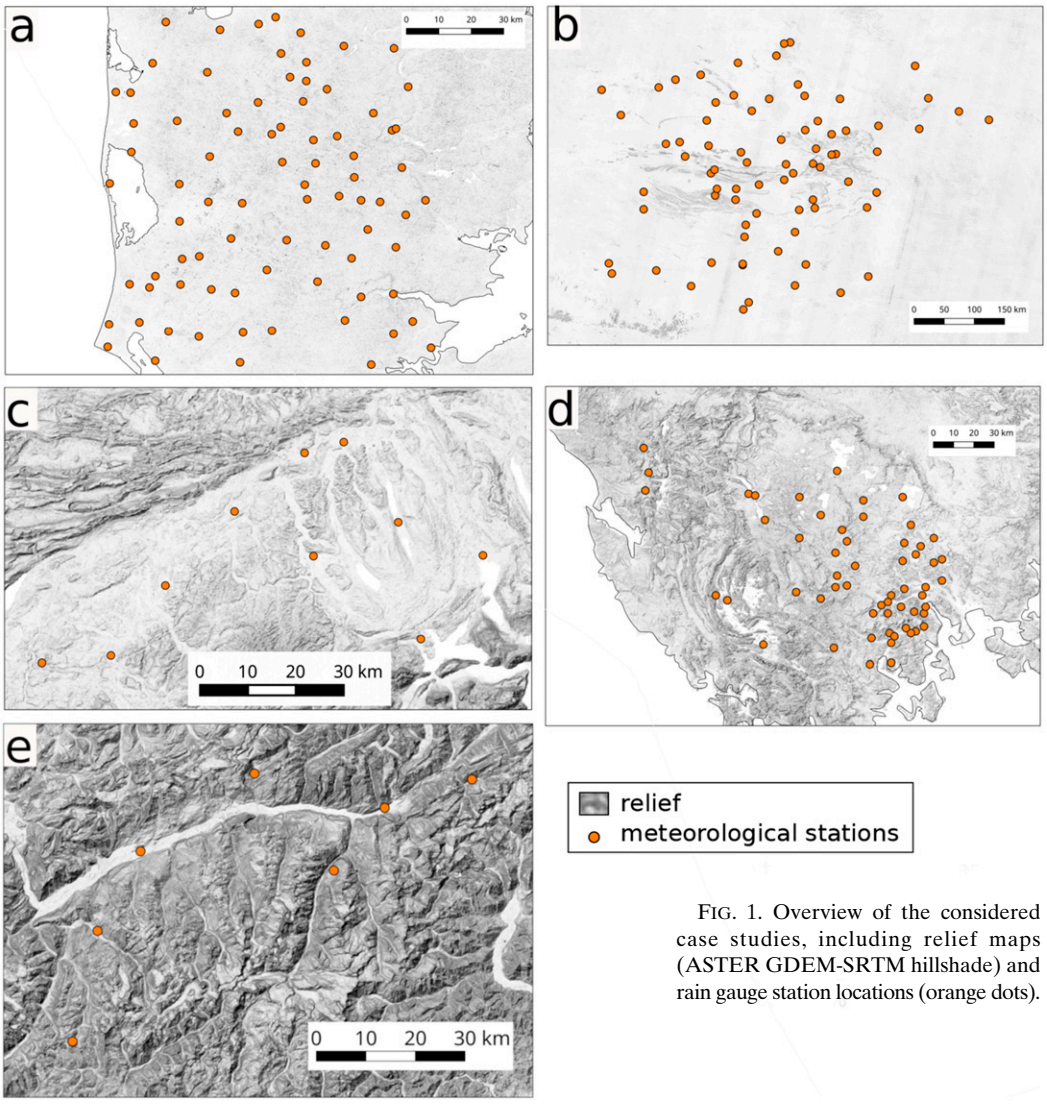

FIG. 1. Overview of the considered case studies, including relief maps (ASTER GDEM-SRTM hillshade) and rain gauge station locations (orange dots).

period, all (natural and artificial) missing data are infilled using different geostatistical interpolation strategies (see section 2), namely, inverse distance weighting (ID), ordinary kriging using a linear variogram model (OKlin), using an exponential model (OKexp), and using indicator kriging to model the dry/wet pattern first (IK + OKlin and IK + OKexp). In addition, IKNN and vector sampling (VS) are proposed as pattern-based strategies.

TABLE 1. Summary of the considered datasets and geographic settings.

Region	Surface (km^2)	Terrain	Predominant climate	Station amount	Temporal coverage	Missing data (%)
Skjern (Denmark)	11 000	Coastal plain	Humid continental	81	1990–2015 (25 years)	30
Alice Springs (Northern Territory)	250 000	Continental plain	Hot desert	76	1987–2019 (33 years)	24
Swiss Plateau (Switzerland)	3000	Hilly alpine foreland	Humid continental	10	1950–2019 (70 years)	31
Southern outback (Tasmania)	16 000	Low mountain to estuary plain	Oceanic	57	1987–2019 (33 years)	33
Wallis (Switzerland)	3400	High mountain range	Cold, no dry season to polar	7	1864–2019 (155 years)	24

After examining a visual example of a filled days–stations data matrix, the results are analyzed with a number of error indicators, computed separately for the four gap scenarios considered. First, the root-mean-square error (RMSE) is computed over all estimated data. To analyze the accuracy in preserving the total rainfall amount over the catchment, the root mean squared error on the sum of the rainfall amounts for each day (sumRMSE) and its bias (sumBIAS) have been computed. To test whether the spatial variability of daily rainfall is preserved, the RMSE and bias have been computed on the standard deviation of the rainfall amount for each day (stdRMSE and stdBIAS). To check whether the dry/wet pattern is preserved, the hit fraction for the dry/wet pattern is computed as $h = (h0 + h1)/2$, where $h0$ corresponds to the fraction of correctly detected dry days (rain amount < 1 mm). Similarly, $h1$ constitutes the ratio of detected wet days. Parameter h varies between 0, for no states correctly detected, and 1, for all states correctly detected.

5. Results

Let us first examine an example of dataset in a days–stations matrix form, containing reference and missing data recovered by the algorithms (Fig. 2). Columns contain data from the same station and rows contain all data from the same day. Stations, represented by columns, are ordered by their easting coordinate. A 100-day portion of validation dataset from the Wallis study case is chosen, that features a distinctive rainfall pattern allowing a clear visual comparison of the different techniques. The reference data (Fig. 2a) show an alternation of dry or low-intensity days (blue-shaded rows, rainfall < 10 mm) and high-intensity days (green-to-yellow-shaded rows, rainfall > 10 mm). Some high-intensity events do not cover all the stations, as shown by the horizontal discontinuities of the yellow rows in the data matrix. Removing approximately 60% of the data (Fig. 2b) leaves the data patterns largely incomplete. As seen in Fig. 2c, ID extends the same intensity observed at informed stations to nearby ones, smoothing out the spatiotemporal pattern. A similar behavior is observed for the kriging-based techniques (Figs. 2d–g). Adding a dry/wet modeling phase based on IK (Figs. 2f,g), does not look to improve the rainfall pattern representation. The rainfall pattern looks to be recovered more realistically by the pattern-based estimation techniques (Figs. 2h,i), but with a limited accuracy for both intense and weak rainfall events.

The following tables show the scores for statistical indicators (see section 4) allowing a quantitative evaluation of the techniques' performance for the different study cases and missing-data scenarios. Table 2 reports the results for the Skjern catchment: the RMSE, representing the average error on each station, is 0.3–0.4 mm lower for interpolation techniques based on ID and OK than the pattern-based ones (IKNN and VS) for a 20% missing-data scenario. A superior performance of these techniques is also shown by the sum of all stations amount over each day (sumRMSE). For both indicators the most accurate technique turns out to be OK using an exponential model (OKexp). Conversely to RMSE, sumRMSE appears to be more sensitive to the lack of data with all techniques, passing from 8–13 to 40–65 mm when the missing data increase from 20% to 80%. This suggests that the overall rainfall amount is more difficult to recover in scarce-data conditions. sumBIAS indicates the tendency to over or underestimate the rainfall amount over the day: ID, OKlin, and OKexp present a sensibly better performance (up to order of magnitude more accurate) than the pattern-based ones, with similar performance and a rather stable behavior when the number of missing data increases. Conversely, IK models, IKNN, and VS tend to underestimate the rainfall amount more and exacerbate this tendency in case of scarce data. The stdRMSE measures the error of the daily rainfall standard deviation, i.e., they inform about the preservation of the spatial rainfall variability. Here, geostatistical interpolation tends to be more accurate, with stdRMSE approximately 0.6 mm lower than pattern-based techniques. The indicator appears to be stable with missing-data increase. The bias on the standard deviation (stdBIAS), shows and overall tendency to underestimate the rainfall variability by all techniques, with a better performance by IK models. Together with sumRMSE, stdBIAS scores appear to be sensitive to the amount of missing data. The last indicator h, informing about the preservation of the dry/wet pattern, shows a rather good performance by all techniques, ranging in 79%–90% of the pattern correctly recovered. ID, OK, and VS tend to be more accurate than IKNN and IK models. In this case, the dry/wet pattern does not look to be significantly affected by the lack of data.

In comparison to the humid region of the Skjern catchment, Alice Springs is again a mainly flat area, but continental and semidesertic: dry conditions and sporadic rainfall events dominate the rainfall heterogeneity. This area is approximately 5 times larger than the previous one, but preserves a similar number of stations, resulting in a lower station density and a possibly larger spatial nonstationarity in the rainfall statistics. The results on this second study case are shown in Table 3. All indicators show a moderately lower performance by all techniques in comparison to the previous study case. Similar tendencies are observed: interpolation techniques,

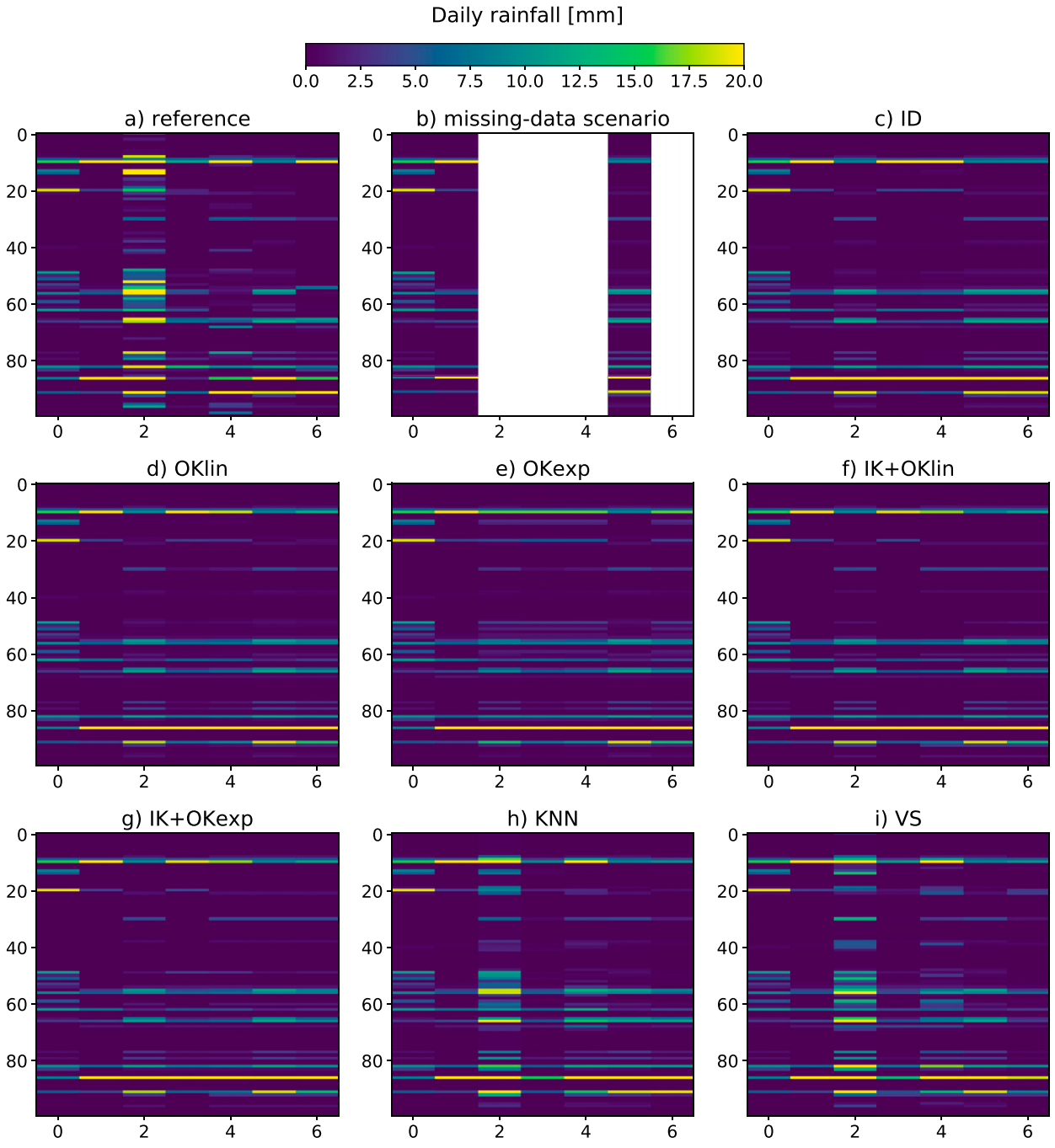

FIG. 2. Example of the gap-filling exercise from a 100-day portion of the daily rainfall record from the Wallis study case. Data are shown in a days–stations matrix form, where each row represents all data in a day and each column all data in a station. The datasets are (a) original data, (b) depleted scenario with 60% of data removed (blank space), and (c)–(i) gap-filled data using the compared algorithms (indicated in the titles).

in particular ID and OKexp, tend to be on average more accurate than pattern-based ones. The dry/wet pattern is more difficult to recover than in the first study case, with h ranging from 0.7% to 0.85% for all techniques.

The third study area, the Swiss Plateau, is a humid region characterized by summer storms and a moderate topographic influence on rainfall nonstationarity. The cross-validation results (Table 4) show a close performance among interpolation and pattern-based techniques,

TABLE 2. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios on the Skjern dataset. The indicator types are RMSE on the daily rainfall amount sum (sumRMSE) and its bias (sumBIAS), RMSE on the daily rainfall standard deviation (stdRMSE) and its bias (stdBIAS), and dry/wet pattern hit fraction *h*. Bold font indicates the best scores.

		Skjern						
Indicator	Missing percent (%)	ID	OKlin	OKexp	IK + OKlin	IK + OKexp	IKNN	VS
RMSE (mm)	20	1.98	1.88	**1.85**	2.04	2.22	2.2	2.23
	40	1.96	1.84	**1.8**	1.98	2.1	2.16	2.15
	60	2.1	2.02	**1.95**	2.13	2.2	2.3	2.21
	80	2.28	2.17	**2.17**	2.33	2.38	2.51	2.34
sumRMSE (mm)	20	8.8	**8.5**	8.93	11.36	17.07	13.06	13.86
	40	14.19	14.1	**13.9**	17.98	28.74	18.88	19.82
	60	24.94	26.01	**23.36**	28.22	39.63	35.69	31.74
	80	40.33	42.07	**40.47**	53.3	60.89	65.84	51.49
sumBIAS (mm)	20	**−0.01**	0.13	0.03	−2.77	−3.45	−1.58	−3.5
	40	**0.5**	0.53	0.68	−5.38	−5.77	−3.13	−5.61
	60	2.7	2.68	**2.38**	−6.52	−7.14	−7.11	−7.28
	80	−0.89	−0.96	**−0.47**	−13.89	−15.13	−17.69	−8.81
stdRMSE (mm)	20	**0.86**	**0.86**	1.12	0.87	0.96	1.63	1.7
	40	0.75	0.75	1.1	**0.74**	0.83	1.51	1.55
	60	**0.76**	0.79	1.17	0.81	0.87	1.6	1.56
	80	**0.92**	1.04	1.62	1.02	1.1	1.72	1.6
stdBIAS (mm)	20	−0.27	−0.29	−0.47	**−0.16**	−0.22	−0.74	−0.84
	40	−0.24	−0.28	−0.5	**−0.15**	−0.21	−0.76	−0.82
	60	−0.29	−0.33	−0.61	**−0.2**	−0.27	−0.83	−0.82
	80	−0.39	−0.48	−0.85	**−0.35**	−0.43	−0.91	−0.82
h (0–1)	20	**0.93**	**0.93**	**0.93**	0.89	0.85	0.76	0.9
	40	**0.93**	**0.93**	0.92	0.89	0.85	0.78	0.9
	60	**0.92**	**0.92**	**0.92**	0.88	0.86	0.79	0.9
	80	0.91	**0.92**	0.91	0.88	0.86	0.81	0.9

except for IKNN that tend to show ≈35% larger RMSE and sumRMSE scores and ≈30% lower dry/wet hit rates (*h*) than to the other techniques. Similarly to previous study cases, sumRMSE results to be the indicator most affected by the missing-data amount.

The fourth study case regards the Tasmanian southern outback, where the influence of topography is critical for the rainfall events, which tend to be more frequent and intense in the western mountains. The results (Table 5) show ≈10%–30% lower RMSE, sumRMSE, and stdRMSE scores by pattern-based techniques in comparison to ID and kriging. Conversely, the sumBIAS, stdBIAS, and *h* scores do not see any predominant technique.

The last study case is the Wallis dataset, where heavy rainfall patterns are influenced by a complex topography. It has to be noted that the rain gauge network is smaller than the other study cases, including only seven stations. Most indicators (Table 6) show the pattern-based techniques, in particular VS, as the best performers, with improvement of 10%–30% on the error and bias scores in comparison to the other methods considered. The dry/wet pattern is recovered less efficiently by IKNN, with *h* scores of ≈60% against ≈85% of the other techniques. The sumRMSE results are again mostly affected by the lack of informed stations.

6. Discussion

The presented tests have shown different behaviors of geostatistical interpolation and pattern-based techniques in estimating missing daily rainfall data. Kriging and inverse distance weighting, being based on spatial averaging, have the tendency to smooth out the spatial rainfall pattern, but also to present a low average error. Conversely, pattern-based techniques such as IKNN and VS apply averaging in time, on historical data patterns that present similar values at informed stations. This type of approach allows recovering more complex local features by bringing information from other days' data. Moreover, the generated structures are possibly extrapolated outside the range of the current-day observations. This makes pattern-based estimation more realistic but it does not necessarily present an optimal fit with the actual rainfall structure, as seen in the visual example of Fig. 2. This results in a tendency to present a higher average error than interpolation techniques. Nevertheless, where rainfall presents a more complex spatial structure, the considered pattern-based techniques allow a more accurate missing-data estimation than spatial interpolation. It has to be noted that IKNN is less effective in recovering the dry/wet pattern than to the other techniques

TABLE 3. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios on the Alice Springs dataset. The indicator types are RMSE on the daily rainfall amount sum (sumRMSE) and its bias (sumBIAS), RMSE on the daily rainfall standard deviation (stdRMSE) and its bias (stdBIAS), and dry/wet pattern hit fraction *h*. Bold font indicates the best scores.

		Alice Springs						
Indicator	Missing percent (%)	ID	OKlin	OKexp	IK + OKlin	IK + OKexp	IKNN	VS
RMSE (mm)	20	3.34	3.06	**2.88**	3.49	4.49	3.02	3.04
	40	2.99	2.82	**2.78**	3.23	3.58	2.95	2.98
	60	3.12	2.98	**2.84**	3.43	3.67	3.19	3.07
	80	3.32	3.28	**3.07**	3.95	3.97	3.45	3.45
sumRMSE (mm)	20	12.35	11.52	**11.06**	16.15	47.9	15.25	15.09
	40	**17.02**	17.86	18.06	27.99	49.42	30.77	25.9
	60	29.46	29.53	**27.38**	49.56	72.24	61.47	37.57
	80	50.57	58.57	**50.39**	85.91	96.04	95.97	59.53
sumBIAS (mm)	20	**0.09**	0.14	0.19	0.88	5.3	−3.09	−2.96
	40	0.39	**0.12**	0.31	1.49	7.85	−7.37	−4.71
	60	−0.68	−1.12	**−0.46**	0.72	6.21	−15.25	−6.57
	80	−0.38	−0.62	**−0.08**	1.67	3.54	−25.15	−4.91
stdRMSE (mm)	20	1.68	**1.56**	2.05	1.85	1.81	1.98	1.89
	40	**1.36**	1.43	2.14	1.52	1.63	1.99	1.83
	60	**1.44**	1.47	2.25	1.63	1.73	2.21	1.8
	80	**1.69**	1.75	2.43	2.0	2.03	2.48	1.85
stdBIAS (mm)	20	−0.14	−0.19	−0.52	0.09	**−0.03**	−0.5	−0.52
	40	−0.2	−0.28	−0.62	**−0.03**	−0.12	−0.57	−0.5
	60	−0.33	−0.38	−0.75	**−0.15**	−0.27	−0.74	−0.52
	80	−0.43	−0.48	−0.85	**−0.26**	−0.38	−0.84	−0.35
h (0–1)	20	0.8	**0.83**	0.85	0.78	0.77	0.8	0.75
	40	0.81	0.83	**0.84**	0.79	0.77	0.77	0.74
	60	0.8	0.81	**0.84**	0.77	0.74	0.73	0.74
	80	0.79	0.8	**0.82**	0.75	0.73	0.66	0.74

considered and VS is usually more precise than IKNN at least for the considered study cases.

With both types of approach, the quality of the estimations depends on the amount of observations at neighbor stations. In particular, the estimated total amount of rainfall turned out to be a sensitive quantity to data availability in the same day. Also, the amount of neighbor data is particularly relevant for kriging techniques that calibrate a variogram model on daily observations, as OKexp used in this study.

Another important aspect of geostatistical interpolation is the choice of the spatial model: for the considered study cases, more parsimonious models such as ID, OKlin, and OKexp present a better performance in comparison to more complex ones such as IK + OKlin and IK + OKexp. IK, whose goal is to model the dry/wet pattern, may require more observations in space to make reliable predictions. An option to improve the calibration if more complex models is to group days into different rainfall types (Hay et al. 1991; Allard and Monestiez 1999; Fowler et al. 2005; Oriani et al. 2017b), but this requires more modeling efforts that are not practical in a routine gap-filling procedure. All simpler models considered here present overall a similar performance: OKlin may be the most convenient approach since it presents no parameters at all to calibrate (see section 2), while OKexp is calibrated on daily data, and ID presents the parameter Θ that is set up empirically (see section 2a). It is worth noting that Θ heavily controls the level of smoothness in the estimation. The value set up in this study ($\Theta = 5$) leads to interpolations dominated by local data and a representation similar to a Voronoi tessellation. A possible reason for a better performance of this setup can be that it allows, in the considered case studies, to better accommodate spatial nonstationarity, preserving sharper transitions. As seen in previous studies (see section 1), this is not always the optimal result and the best Θ value can vary with the considered dataset.

Conversely, IKNN and VS present the parameter K that controls the number of best-matching patterns used for the estimation. This needs to be set up for specific case studies, usually taking values in the range 5–20 (see sections 2c and 2d), with the goal of including a sufficiently large set of days representing the same type of events as the target one. There is currently no clear strategy to setup K a priori, the parameter being sensitive to the complexity of the historical data and the variability of daily rainfall events. Nevertheless, VS is sufficiently fast to allow setting K empirically with a

TABLE 4. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios on the Swiss Plateau dataset. The indicator types are RMSE on the daily rainfall amount sum (sumRMSE) and its bias (sumBIAS), RMSE on the daily rainfall standard deviation (stdRMSE) and its bias (stdBIAS), and dry/wet pattern hit fraction *h*. Bold font indicates the best scores.

		Swiss Plateau						
Indicator	Missing percent (%)	ID	OKlin	OKexp	IK + OKlin	IK + OKexp	IKNN	VS
RMSE (mm)	20	3.26	**3.09**	3.18	3.18	3.19	4.47	3.15
	40	3.15	**3.04**	3.2	3.16	3.21	4.33	3.33
	60	3.14	**3.07**	3.14	3.16	3.22	4.42	3.39
	80	3.85	3.78	**3.75**	3.86	3.86	5.1	3.99
sumRMSE (mm)	20	5.79	5.56	**5.54**	5.77	5.82	7.81	5.67
	40	**7.3**	7.32	7.8	7.75	8.15	13.52	9.1
	60	11.2	**11.17**	11.19	11.31	12.1	18.76	12.57
	80	20.58	20.53	**20.1**	20.84	20.96	32.13	20.78
sumBIAS (mm)	20	0.13	0.17	0.17	−0.28	−0.27	−0.09	**−0.03**
	40	−0.13	−0.1	−0.18	−0.92	−1.06	**0.05**	−0.42
	60	0.26	0.28	0.27	−0.93	−1.16	0.49	**0.05**
	80	−0.48	**−0.45**	−0.53	−1.3	−1.37	−0.73	−0.81
stdRMSE (mm)	20	1.56	**1.53**	1.89	**1.53**	1.54	1.74	1.57
	40	**1.55**	1.6	2.25	1.61	1.65	1.88	1.71
	60	**1.67**	1.76	2.38	1.77	1.83	1.89	1.75
	80	2.23	2.29	2.73	2.29	2.3	2.08	**1.92**
stdBIAS (mm)	20	−0.3	−0.38	−0.63	−0.37	−0.38	**0.2**	−0.35
	40	**−0.39**	−0.49	−0.9	−0.41	−0.47	−0.05	−0.45
	60	−0.51	−0.67	−1.11	−0.59	−0.66	**−0.01**	−0.47
	80	−0.91	−1.01	−1.33	−0.96	−1.0	**−0.08**	−0.57
h (0–1)	20	**0.92**	**0.92**	**0.92**	0.9	0.88	0.6	0.91
	40	**0.91**	**0.91**	**0.91**	0.88	0.87	0.6	**0.91**
	60	**0.91**	**0.91**	**0.91**	0.89	0.88	0.58	0.9
	80	**0.89**	**0.89**	0.88	0.88	0.88	0.51	0.88

series of cross-validation tests. Once done on a dataset including different years, the setup should be valid for subsequent daily updates. We recommend using the longest dataset available for the study region, including days presenting missing data, which are handled robustly by both VS and IKNN. Also, the quality of the estimation is supposed to improve by updating the training dataset with new data.

Computational time is comparable among all tested interpolation models and it is generally two orders of magnitude faster than the VS and IKNN implementations used (section 2), that are not currently optimized for speed. IKNN is fully sequential on each station and tends to be slower than VS, which only iterates over days and simulates together the main part of missing data patterns. The computation time remains on the order of minutes for thousands of missing data. Both workflows can be optimized for parallel computing.

To summarize, the results of this study suggest that interpolation models are a more likely a reliable option in case of flat terrains and mainly homogeneous rainfall events, e.g., the study cases of Skjern considered in this study. In particular, simpler models seem to be more reliable than the complex ones when a limited quantity of data is available in space. On the other hand, pattern-based approaches such as VS, preferable over IKNN for its higher accuracy, can be a valid alternative to geostatistical interpolation techniques when the study zone presents a more complex rainfall heterogeneity and nonstationarity driven by external factors like elevation. Two examples in this study are the Tasmania outback and the Wallis region.

7. Conclusions

In this study, we have analyzed the potential of geostatistical interpolation and data-pattern estimation to generate missing daily rain gauge data. The goal was to compare techniques that are not overly demanding in terms of modeling efforts and can be used for automatic gap filling.

The considered interpolation techniques include the popular inverse distance weighting and different models based on ordinary kriging. As pattern-based approaches, we have proposed an iterative version of the *K*-nearest neighbor, and a novel algorithm called vector sampling (VS). Both techniques base the estimation of missing data on historical data patterns similar to the one constituted of informed stations in the target day. Their

TABLE 5. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios in the southern outback of Tasmania. The indicator types are RMSE on the daily rainfall amount sum (sumRMSE) and its bias (sumBIAS), RMSE on the daily rainfall standard deviation (stdRMSE) and its bias (stdBIAS), and dry/wet pattern hit fraction h. Bold font indicates the best scores.

		Tasmanian southern outback						
Indicator	Missing percent (%)	ID	OKlin	OKexp	IK + OKlin	IK + OKexp	IKNN	VS
RMSE (mm)	20	6.81	6.46	6.08	7.05	7.41	**4.83**	4.88
	40	6.65	6.33	6.14	6.53	6.94	**4.98**	5.03
	60	6.43	6.02	6.17	6.13	6.46	5.31	**5.19**
	80	6.65	6.53	6.54	7.1	7.32	6.16	**5.69**
sumRMSE (mm)	20	23.05	21.91	20.97	27.47	33.3	**16.61**	17.09
	40	33.32	34.64	33.06	39.84	57.55	31.08	**30.42**
	60	51.37	46.25	53.69	51.59	68.47	54.23	**45.14**
	80	105.43	107.02	114.01	119.42	137.23	121.47	**75.33**
sumBIAS (mm)	20	**−0.66**	1.24	1.02	−0.55	−2.14	−2.7	−2.94
	40	−3.05	**−1.02**	−1.41	−5.97	−9.51	−10.24	−7.93
	60	−12.36	**−7.05**	−7.89	−14.74	−18.88	−16.38	−9.2
	80	−22.15	−15.31	−13.12	−25.61	−26.84	−36.37	**−6.88**
stdRMSE (mm)	20	5.53	5.19	5.06	5.59	5.75	3.82	**3.81**
	40	5.06	4.83	4.98	4.97	5.16	3.84	**3.8**
	60	4.81	4.54	5.12	4.55	4.7	4.06	**3.8**
	80	4.75	4.75	5.24	5.13	5.24	4.31	**3.98**
stdBIAS (mm)	20	−0.2	−0.13	−0.71	**0.09**	−0.23	−0.78	−0.67
	40	−0.54	−0.67	−1.32	**−0.48**	−0.83	−1.37	−0.98
	60	−1.62	−1.51	−2.3	−1.26	−1.56	−1.36	**−0.93**
	80	−2.14	−2.16	−2.99	−1.99	−2.25	−1.38	**−0.92**
h (0–1)	20	**0.89**	0.88	0.88	0.85	0.79	0.69	0.85
	40	**0.88**	**0.88**	**0.88**	0.84	0.78	0.7	0.86
	60	**0.87**	**0.87**	0.86	0.83	0.79	0.67	0.85
	80	0.84	0.84	0.84	0.82	0.79	0.62	**0.85**

peculiarity is their capacity to handle multiple missing data in both the predictor and training variables. This is a strategic feature, considering that in real practice the dataset to complete has often to be used as training dataset as well.

The presented tests include five case studies featuring progressively complex terrains and rainfall heterogeneity. The results do not suggest a relation to specific climate settings, but rather to the level of complexity of the associated rainfall patterns. Simple interpolation is more likely the best option in case of flat terrains and mainly homogeneous rainfall events, while on complex terrains and rainfall events the proposed pattern-based techniques, in particular VS, turns out to be more accurate. In its early implementation, VS requires more computation time than geostatistical interpolation, but stays on the order of a few minutes for thousands of missing data, with no separate training phase required. As usual for pattern-based techniques, VS performs best when trained by a representative historical dataset that can be the target dataset itself. For its robustness in handling sparse missing data and its minimal parameterization, the technique can be easily incorporated into automatic dataset maintenance by practitioners and public institutions.

This study poses new questions and possible developments on missing-data generation: what is the right balance between minimization of the average error and preservation of complex spatial features? What are the characteristics of rainfall that are more important to preserve in relation to physical models that use rainfall as input? Further testing is necessary, in particular considering the hydrological response of regional domains where the spatial rainfall distribution can be highly nonstationary.

8. Computer code availability

Vector Sampling is freely available as open-source code:

- Name: VS
- Language: Python 3
- Version: 1.0
- Year developed: 2019
- Developed by: Fabio Oriani at the Geological Survey of Denmark and Greenland (GEUS) and the University of Lausanne (UNIL)

TABLE 6. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios on the Wallis dataset. The indicator types are RMSE on the daily rainfall amount sum (sumRMSE) and its bias (sumBIAS), RMSE on the daily rainfall standard deviation (stdRMSE) and its bias (stdBIAS), and dry/wet pattern hit fraction *h*. Bold font indicates the best scores.

		Wallis						
Indicator	Missing percent (%)	ID	OKlin	OKexp	IK + OKlin	IK + OKexp	IKNN	VS
RMSE (mm)	20	5.66	5.69	5.47	5.75	5.79	4.73	**4.4**
	40	5.54	5.53	5.37	5.61	5.68	4.74	**4.65**
	60	5.7	5.53	5.53	5.54	5.73	5.23	**4.93**
	80	6.97	6.93	6.96	6.99	6.99	6.29	**5.44**
sumRMSE (mm)	20	7.49	7.64	7.19	7.7	7.74	6.11	**5.46**
	40	12.6	12.65	12.48	12.65	12.86	10.02	**9.1**
	60	15.29	15.45	15.98	15.42	16.43	14.83	**12.74**
	80	31.06	31.04	31.05	31.22	31.22	26.38	**19.78**
sumBIAS (mm)	20	−1.37	−1.16	**−0.18**	−1.55	−1.52	−0.33	−0.21
	40	−0.86	−0.59	0.51	−1.12	−1.11	**−0.16**	**−0.16**
	60	−1.81	−1.27	−0.35	−2.94	−2.68	1.09	**−0.31**
	80	2.41	2.43	2.46	2.11	2.12	3.16	**0.62**
stdRMSE (mm)	20	2.66	2.61	2.78	2.67	2.65	2.02	**1.97**
	40	2.38	2.24	2.66	2.39	2.44	**2.08**	2.12
	60	3.26	3.15	3.8	3.25	3.28	2.82	**2.77**
	80	3.95	4.02	4.37	3.99	4.0	3.29	**3.07**
stdBIAS (mm)	20	−0.59	−0.57	−0.77	−0.57	−0.6	**−0.03**	−0.12
	40	−0.53	−0.55	−0.92	−0.43	−0.54	**−0.01**	−0.19
	60	−0.89	−1.17	−1.65	−1.15	−1.23	**0.06**	−0.37
	80	−1.92	−2.0	−2.23	−1.96	−1.96	0.33	**−0.27**
h (0–1)	20	0.82	0.81	0.83	0.8	0.79	0.64	**0.88**
	40	0.83	0.81	0.83	0.8	0.79	0.59	**0.86**
	60	0.81	0.81	0.81	0.8	0.79	0.51	**0.84**
	80	0.8	0.8	**0.81**	0.8	0.8	0.5	0.78

- Repository (code and tutorials): https://bitbucket.org/orianif/vs

Acknowledgments. This research has been funded by the Swiss National Science Foundation (project P2NEP2_162040) and hosted by the SPACE project (http://space.geus.dk) and the GAIA Lab (http://wp.unil.ch/gaia). The data used are available from the HOBE project (http://www.hobe.dk). We acknowledge the financial support for the SPACE project by the Villum Foundation (http://villumfonden.dk/) through their Young Investigator Programme (Grant VKR023443). The third author (MCD) is supported by the National Center for High Performance Computing of Turkey (UHeM) under Grant 1007292019.

APPENDIX A

Pseudocode for the Vector Sampling Algorithm

Input: training dataset ($\mathbf{R}_T$), dataset to complete ($\mathbf{R}_{\text{out}}$), and parameter K. $\mathbf{R}_T$ and $\mathbf{R}_{\text{out}}$ are days × stations data matrices and must have the same number of columns (stations) or be the same dataset. K is the number of candidate days used for the estimation.

1. **While** missing data exist in $\mathbf{R}_{\text{out}}$ **do**:
   2. **for** each day $\mathbf{r}_{\text{out}}$ in $\mathbf{R}_{\text{out}}$ **do**:
      3. **if** $\mathbf{r}_{\text{out}}$ contains missing data **do**:
         4. compute the pattern distance **D** between $\mathbf{r}_{\text{out}}$ and all days in $\mathbf{R}_T$;
         5. for missing data in $\mathbf{R}_T$ assign in **D** the mean distance of the station;
         6. from **D** compute the mean distance for each day $\mathbf{d}_i$;
         7. impose $\mathbf{d}_i = +\text{Inf}$ for days that do not add any data to $\mathbf{r}_{\text{out}}$;
         8. from $\mathbf{R}_T$ select the K days $\mathbf{R}_K$ showing the lowest $\mathbf{d}_i$;
         9. estimate the missing data in $\mathbf{r}_{\text{out}}$ as the weighted mean of $\mathbf{R}_K$ using as weights $1/\mathbf{d}_i$.

Output: the completed dataset $\mathbf{R}_{\text{out}}$.

APPENDIX B

Imposed Pattern-Mismatch Error at Uninformed Locations

To test the efficiency of the distance estimator for incomplete data patterns [Eq. (10)], we perform a

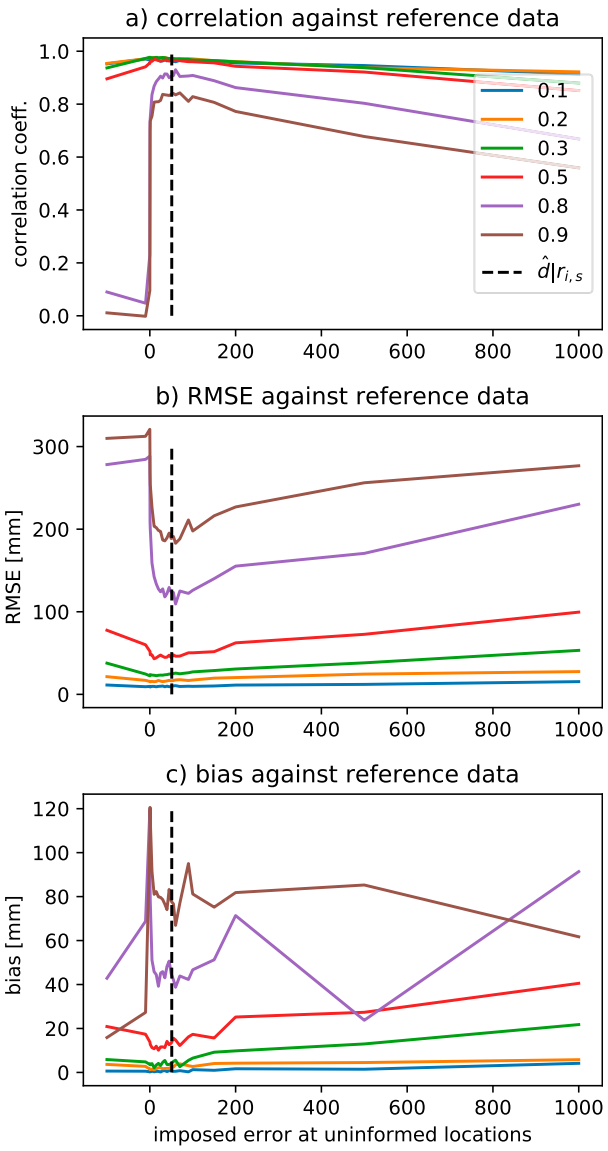

FIG. B1. Performance indexes as a function of the imposed error at uninformed locations. Different random missing-data fractions are considered as shown in the color legend. The dashed line indicates the value estimated with Eq. (B1).

cross-validation experiment where different random data amounts are removed from a portion of the Skjern historical rainfall dataset, containing 5000 days and 80 stations. This dataset is used as target and training dataset at the same time. For all missing-data scenarios, the dataset is reconstructed using algorithm (8), but instead of computing the error estimator for missing data using Eq. (10), we test a series of arbitrary values as constant imposed error at missing locations. We then observe how the performance of the reconstruction varies as a function of this parameter. As performance indicators, we use the Pearson's correlation coefficient, the RMSE, and the bias between the simulated and reference data artificially removed. Then the results are compared with the average error estimator of the entire dataset. To do that, we extend the conditional estimator of Eq. (10) to the entire database by computing the average $\hat{d}$ for all informed i days and s stations:

$$\hat{d}|r_{i,s} = \left(\sum_{s=1}^{S}\sum_{i=1}^{I}\mathbb{I}_{r_{i,s}\neq\mathrm{NaN}}\right)^{-1}\sum_{s=1}^{S}\sum_{\substack{i=1,\\ r_{i,s}\neq\mathrm{NaN}}}^{I}\hat{d}_s|r_{i,s}. \qquad \text{(B1)}$$

Figure B1 shows that, for conspicuous missing-data amounts, the quality of the estimation is heavily dependent on the imposed error for missing data in terms of correlation with the reference (Fig. B1a), mean absolute error (Fig. B1b), and bias (Fig. B1c). Extreme values of the imposed error lead to a poor quality in the estimation. In particular, too low values lead to use as a data source for very fragmented patterns that may not have a high prediction power on the missing data. Conversely, too high error values lead us to discard any incomplete pattern and choose only the complete ones, which may constitute an insufficient data source. In this case, the optimal balance is achieved with a value close to 50, which matches well with $\hat{d}|r_{i,s}$. This result suggest that the conditional absolute error is a robust metric to estimate the distance at uninformed locations and allow sampling from incomplete patterns.

REFERENCES

Abebe, A. J., D. P. Solomatine, and R. G. W. Venneker, 2000: Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. *Hydrol. Sci. J.*, **45**, 425–436, https://doi.org/10.1080/02626660009492339.

Allard, D., and P. Monestiez, 1999: Geostatistical segmentation of rainfall data. *geoENV II — Geostatistics for Environmental Applications*, J. Gómez-Hernández, A. Soares, and R. Froidevaux, Eds., Quantitative Geology and Geostatistics Series, Vol. 10, Springer, 139–150, https://doi.org/10.1007/978-94-015-9297-0_12.

Apipattanavis, S., G. Podesta, B. Rajagopalan, and R. W. Katz, 2007: A semiparametric multivariate and multisite weather generator. *Water Resour. Res.*, **43**, W11401, https://doi.org/10.1029/2006WR005714.

ASCE, 1996: *Hydrology Handbook*. ASCE, 784 pp.

Azman, M. A.-Z., R. Zakaria, and N. F. A. Radi, 2015: Estimation of missing rainfall data in Pahang using modified spatial interpolation weighting methods. *AIP Conf. Proc.*, **1643**, 65–72, https://doi.org/10.1063/1.4907426.

Bárdossy, A., and G. Pegram, 2014: Infilling missing precipitation records - A comparison of a new copula-based method with other techniques. *J. Hydrol.*, **519**, 1162–1170, https://doi.org/10.1016/j.jhydrol.2014.08.025.

Bayabil, H. K., A. Fares, H. O. Sharif, D. T. Ghebreyesus, and H. A. Moreno, 2019: Effects of spatial and temporal data

aggregation on the performance of the multi-radar multi-sensor system. *J. Amer. Water Resour. Assoc.*, **55**, 1492–1504, https://doi.org/10.1111/1752-1688.12799.

Benoit, L., D. Allard, and G. Mariethoz, 2018a: Stochastic rainfall modeling at sub-kilometer scale. *Water Resour. Res.*, **54**, 4108–4130, https://doi.org/10.1029/2018WR022817.

——, M. Vrac, and G. Mariethoz, 2018b: Dealing with non-stationarity in sub-daily stochastic rainfall models. *Hydrol. Earth Syst. Sci.*, **22**, 5919–5933, https://doi.org/10.5194/hess-22-5919-2018.

Burhanuddin, S. N. Z. A., S. M. Deni, and N. M. Ramli, 2015: Geometric median for missing rainfall data imputation. *AIP Conf. Proc.*, **1643**, 113–119, https://doi.org/10.1063/1.4907433.

Caldera, H. P. G. M., V. R. P. C. Piyathisse, and K. D. W. Nandalal, 2016: A comparison of methods of estimating missing daily rainfall data. *Eng.: J. Inst. Eng., Sri Lanka*, **49** (4), 1–8, https://doi.org/10.4038/engineer.v49i4.7232.

Caraway, N. M., J. L. McCreight, and B. Rajagopalan, 2014: Multisite stochastic weather generation using cluster analysis and k-nearest neighbor time series resampling. *J. Hydrol.*, **508**, 197–213, https://doi.org/10.1016/j.jhydrol.2013.10.054.

Chappell, A., L. J. Renzullo, T. H. Raupach, and M. Haylock, 2013: Evaluating geostatistical methods of blending satellite and gauge data to estimate near real-time daily rainfall for Australia. *J. Hydrol.*, **493**, 105–114, https://doi.org/10.1016/j.jhydrol.2013.04.024.

Cooke, R., and S. Mostaghimi, 1992: A microcomputer-based routine for obtaining mean watershed precipitation from point values. *Comput. Geosci.*, **18**, 823–837, https://doi.org/10.1016/0098-3004(92)90027-O.

Cressie, N., 1985: Fitting variogram models by weighted least squares. *J. Int. Assoc. Math. Geol.*, **17**, 563–586, https://doi.org/10.1007/BF01032109.

Dembélé, M., F. Oriani, J. Tumbulto, G. Mariéthoz, and B. Schaefli, 2019: Gap-filling of daily streamflow time series using direct sampling in various hydroclimatic settings. *J. Hydrol.*, **569**, 573–586, https://doi.org/10.1016/j.jhydrol.2018.11.076.

Demirel, M. C., J. Mai, G. Mendiguren, J. Koch, L. Samaniego, and S. Stisen, 2018: Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrol. Earth Syst. Sci.*, **22**, 1299–1315, https://doi.org/10.5194/hess-22-1299-2018.

Di Piazza, A., F. Lo Conti, L. V. Noto, F. Viola, and G. La Loggia, 2011: Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. *Int. J. Appl. Earth Obs. Geoinf.*, **13**, 396–408, https://doi.org/10.1016/j.jag.2011.01.005.

Fowler, H., C. Kilsby, P. O'connell, and A. Burton, 2005: A weather-type conditioned multi-site stochastic rainfall model for the generation of scenarios of climatic variability and change. *J. Hydrol.*, **308**, 50–66, https://doi.org/10.1016/j.jhydrol.2004.10.021.

Grayson, R. B., G. Blöschl, A. W. Western, and T. A. McMahon, 2002: Advances in the use of observed spatial patterns of catchment hydrological response. *Adv. Water Resour.*, **25**, 1313–1334, https://doi.org/10.1016/S0309-1708(02)00060-X.

Guilloteau, C., R. Roca, and M. Gosset, 2016: A multiscale evaluation of the detection capabilities of high-resolution satellite precipitation products in West Africa. *J. Hydrometeor.*, **17**, 2041–2059, https://doi.org/10.1175/JHM-D-15-0148.1.

Hay, L. E., G. J. Mccabe, D. M. Wolock, and M. A. Ayers, 1991: Simulation of precipitation by weather type analysis. *Water Resour. Res.*, **27**, 493–501, https://doi.org/10.1029/90WR02650.

He, X., J. Koch, C. Zheng, T. Bøvith, and K. H. Jensen, 2018: Comparison of simulated spatial patterns using rain gauge and polarimetric-radar-based precipitation data in catchment hydrological modeling. *J. Hydrometeor.*, **19**, 1273–1288, https://doi.org/10.1175/JHM-D-17-0235.1.

Hema, N., and K. Kant, 2017: Reconstructing missing hourly real-time precipitation data using a novel intermittent sliding window period technique for automatic weather station data. *J. Meteor. Res.*, **31**, 774–790, https://doi.org/10.1007/s13351-017-6084-8.

Hu, J., J. Liu, Y. Liu, and C. Gao, 2013: EMD-KNN model for annual average rainfall forecasting. *J. Hydrol. Eng.*, **18**, 1450–1457, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000481.

Huffman, G. J., R. F. Adler, B. Rudolf, U. Schneider, and P. R. Keehn, 1995: Global precipitation estimates based on a technique for combining satellite-based estimates, rain gauge analysis, and NWP model precipitation information. *J. Climate*, **8**, 1284–1295, https://doi.org/10.1175/1520-0442(1995)008<1284:GPEBOA>2.0.CO;2.

Ismail, W. N. W., and W. Z. W. Z. W. Ibrahim, 2017: Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods. *Malays. J. Fundam. Appl. Sci.*, **13**, 214–218, https://doi.org/10.11113/MJFAS.V13N3.578.

Kajornrit, J., K. W. Wong, and C. C. Fung, 2012a: Estimation of missing precipitation records using modular artificial neural networks. *Neural Information Processing: Lecture Notes in Computer Science*, T. W. Huang et al., Eds., Springer, 52–59.

——, ——, and ——, 2012b: Rainfall prediction in the northeast region of Thailand using modular fuzzy inference system. *2012 IEEE Int. Conf. on Fuzzy Systems*, Brisbane, QLD, Australia, IEEE, 1–6, https://doi.org/10.1109/FUZZ-IEEE.2012.6250785.

Kim, J.-W., and Y. A. Pachepsky, 2010: Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *J. Hydrol.*, **394**, 305–314, https://doi.org/10.1016/j.jhydrol.2010.09.005.

Kisaka, M. O., M. Mucheru-Muna, F. K. Ngetich, J. Mugwe, D. Mugendi, F. Mairura, C. Shisanya, and G. L. Makokha, 2016: Potential of deterministic and geostatistical rainfall interpolation under high rainfall variability and dry spells: Case of Kenya's Central Highlands. *Theor. Appl. Climatol.*, **124**, 349–364, https://doi.org/10.1007/S00704-015-1413-2.

Koch, J., G. Mendiguren, G. Mariethoz, and S. Stisen, 2017: Spatial sensitivity analysis of simulated land surface patterns in a catchment model using a set of innovative spatial performance metrics. *J. Hydrol.*, **18**, 1121–1142, https://doi.org/10.1175/JHM-D-16-0148.1.

Lasser, M., O. Sungmin, and U. Foelsche, 2019: Evaluation of GPM-DPR precipitation estimates with WegenerNet gauge data. *Atmos. Meas. Tech.*, **12**, 5055–5070, https://doi.org/10.5194/amt-12-5055-2019.

Linsley, R. K., M. A. Kohler, and J. Paulhus, 1988: *Hydrology for Engineers*. McGraw-Hill, 492 pp.

Mair, A., and A. Fares, 2010: Assessing rainfall data homogeneity and estimating missing records in Mamacrkaha Valley, O'ahu, Hawaii. *J. Hydrol. Eng.*, **15**, 61–66, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000145.

Mariethoz, G., P. Renard, and J. Straubhaar, 2010: The direct sampling method to perform multiple-point geostatistical simulations. *Water Resour. Res.*, **46**, W11536, https://doi.org/10.1029/2008WR007621.

Matheron, G., 1963: Principles of geostatistics. *Econ. Geol.*, **58**, 1246–1266, https://doi.org/10.2113/gsecongeo.58.8.1246.

Moeletsi, M. E., Z. P. Shabalala, G. De Nysschen, and S. Walker, 2016: Evaluation of an inverse distance weighting method for patching daily and dekadal rainfall over the Free State Province, South Africa. *Water S.A.*, **42**, 466–474, https://doi.org/10.4314/wsa.v42i3.12.

New, M., M. Todd, M. Hulme, and P. Jones, 2001: Precipitation measurements and trends in the twentieth century. *Int. J. Climatol.*, **21**, 1889–1922, https://doi.org/10.1002/joc.680.

Noori, M. J., H. H. Hassan, and Y. T. Mustafa, 2014: Spatial estimation of rainfall distribution and its classification in Duhok Governorate using GIS. *J. Water Resour. Prot.*, **06**, 75–82, https://doi.org/10.4236/jwarp.2014.62012.

Oriani, F., J. Straubhaar, P. Renard, and G. Mariethoz, 2014: Simulation of rainfall time series from different climatic regions using the direct sampling technique. *Hydrol. Earth Syst. Sci.*, **18**, 3015–3031, https://doi.org/10.5194/hess-18-3015-2014.

——, A. Borghi, J. Straubhaar, G. Mariethoz, and P. Renard, 2016: Missing data simulation inside flow rate time-series using multiple-point statistics. *Environ. Modell. Software*, **86**, 264–276, https://doi.org/10.1016/j.envsoft.2016.10.002.

——, R. Mehrotra, G. Mariethoz, J. Straubhaar, A. Sharma, and P. Renard, 2017a: Simulating rainfall time-series: How to account for statistical variability at multiple scales? *Stochastic Environ. Res. Risk Assess.*, **32**, 321–340, https://doi.org/10.1007/s00477-017-1414-z.

——, N. Ohana-Levi, F. Marra, J. Straubhaar, G. Mariethoz, P. Renard, A. Karnieli, and E. Morin, 2017b: Simulating small-scale rainfall fields conditioned by weather state and elevation: A data-driven approach based on rainfall radar images. *Water Resour. Res.*, **53**, 8512–8532, https://doi.org/10.1002/2017WR020876.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Rubin, D. B., 2004: *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 258 pp.

Seo, B.-C., and Coauthors, 2018: Comprehensive evaluation of the IFloodS radar rainfall products for hydrologic applications. *J. Hydrometeor.*, **19**, 1793–1813, https://doi.org/10.1175/JHM-D-18-0080.1.

Stisen, S., and M. Tumbo, 2015: Interpolation of daily rain-gauge data for hydrological modelling in data sparse regions using pattern information from satellite data. *Hydrol. Sci. J.*, **60**, 1911–1926, https://doi.org/10.1080/02626667.2014.992789.

Stone, C. J., 1977: Consistent nonparametric regression. *Ann. Stat.*, **5**, 595–620, https://doi.org/10.1214/aos/1176343886.

Strebelle, S., 2002: Conditional simulation of complex geological structures using multiple-point statistics. *Math. Geol.*, **34** (1), 1–21, https://doi.org/10.1023/A:1014009426274.

Teegavarapu, R. S. V., 2012: Spatial interpolation using nonlinear mathematical programming models for estimation of missing precipitation records. *Hydrol. Sci. J.*, **57**, 383–406, https://doi.org/10.1080/02626667.2012.665994.

——, and V. Chandramouli, 2005: Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.*, **312**, 191–206, https://doi.org/10.1016/j.jhydrol.2005.02.015.

Tung, Y.-K., 2013: Evaluation of point rainfall estimation methods in Hong Kong. *35th World Congress of the Int. Association for Hydro-Environment-Engineering and Research*, Chengdu, China, IAHR, 453.

van Buuren, S., and K. Groothuis-Oudshoorn, 2010: mice: Multivariate imputation by chained equations in R. *J. Stat. Software*, **45**, 1–68, https://doi.org/10.18637/JSS.V045.I03.

Verworn, A., and U. Haberlandt, 2011: Spatial interpolation of hourly rainfall - Effect of additional information, variogram inference and storm properties. *Hydrol. Earth Syst. Sci.*, **15**, 569–584, https://doi.org/10.5194/hess-15-569-2011.

Wagner, P. D., P. Fiener, F. Wilken, S. Kumar, and K. Schneider, 2012: Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. *J. Hydrol.*, **464–465**, 388–400, https://doi.org/10.1016/j.jhydrol.2012.07.026.

Wei, T. C., and J. L. McGuinness, 1973: *Reciprocal Distance Squared Method, a Computer Technique for Estimating Areal Precipitation*. Agricultural Research Service, 29 pp.

Woldesenbet, T. A., N. A. Elagib, L. Ribbe, and J. Heinrich, 2017: Gap filling and homogenization of climatological datasets in the headwater region of the Upper Blue Nile Basin, Ethiopia. *Int. J. Climatol.*, **37**, 2122–2140, https://doi.org/10.1002/joc.4839.

Wu, J. S., 2009: A novel artificial neural network ensemble model based on k-nearest neighbor nonparametric estimation of regression function and its application for rainfall forecasting. *2009 Int. Joint Conf. on Computational Sciences and Optimization*, Sanya, China, 41–53, https://doi.org/10.1109/CSO.2009.307.

Yang, X., X. Xie, D. L. Liu, F. Ji, and L. Wang, 2015: Spatial interpolation of daily rainfall data for local climate impact assessment over greater Sydney region. *Adv. Meteor.*, **2012**, 563629, https://doi.org/10.1155/2015/563629.