Estimating extremely large amounts of missing precipitation data

Héctor Aguilera, Carolina Guardiola-Albert and Carmen Serrano-Hidalgo

ABSTRACT

Accurate estimation of missing daily precipitation data remains a difficult task. A wide variety of methods exists for infilling missing values, but the percentage of gaps is one of the main factors limiting their applicability. The present study compares three techniques for filling in large amounts of missing daily precipitation data: spatio-temporal kriging (STK), multiple imputation by chained equations through predictive mean matching (PMM), and the random forest (RF) machine learning algorithm. To our knowledge, this is the first time that extreme missingness (>90%) has been considered. Different percentages of missing data and missing patterns are tested in a large dataset drawn from 112 rain gauges in the period 1975–2017. The results show that both STK and RF can handle extreme missingness, while PMM requires larger observed sample sizes. STK is the most robust method, suitable for chronological missing patterns. RF is efficient under random missing patterns. Model evaluation is usually based on performance and error measures. However, this study outlines the risk of just relying on these measures without checking for consistency. The RF algorithm overestimated daily precipitation outside the validation period in some cases due to the overdetection of rainy days under time-dependent missing patterns.

Key words | evaluation, large missing precipitation, multiple imputation, random forest, spatio-temporal kriging

Héctor Aguilera (corresponding author)
Carolina Guardiola-Albert
Carmen Serrano-Hidalgo
Research on Geological Resources,
Geological Survey of Spain,
Ríos Rosas 23, 28003, Madrid,
Spain
E-mail: *h.aguilera@igme.es*

Carmen Serrano-Hidalgo
School of Mining Engineering of Madrid,
Technical University of Madrid,
Ríos Rosas 21, 28003, Madrid,
Spain

INTRODUCTION

Accurate estimation of missing precipitation data remains a difficult task, particularly for large watersheds with sparse rain gauge networks and large numbers of missing values (MV). The high spatio-temporal variability of precipitation makes it a difficult variable to deal with. Representative precipitation time series are essential to develop consistent hydrological or hydrogeological models for suitable water management (Nkiaka *et al.* 2016; Ben Aissia *et al.* 2017).

Problems with missing data in climatic series often arise and are caused by many circumstances, mainly due to the sources of acquisition, which are usually reports, manual collection instruments, or remote sensors. Typically, these problems lead to a combination of random and chronological missing data patterns in precipitation time series.

The problem of MV in meteorological series is particularly significant in developing countries where gauging stations are scarce and the degree of missingness is large (Yozgatligil *et al.* 2013; Radi *et al.* 2015; Nkiaka *et al.* 2016). However, the issue becomes global when long series (>30 years) or remote watersheds are considered (Ben Aissia *et al.* 2017).

Simply ignoring missing data can lead to partial and biased results in data analysis (Harel & Zhou 2007). A wide variety of methods exists for infilling MV, but the percentage of gaps is one of the main factors limiting their applicability (Lo Presti *et al.* 2010; Yozgatligil *et al.* 2013; Miró *et al.* 2017). Simple methods such as mean imputation and linear interpolation (which just rely on the available

doi: 10.2166/hydro.2020.127

information of the time series to be completed), arithmetic averaging, weighted averaging (usually referred to as the normal ratio method) and inverse distance weighting with data from neighboring stations, have shown poor performance when the amount of MV is large (>5–10% MV; Johnson 2003 in Lo Presti *et al.* 2010). In recent years, progressively more advanced methods have been applied to fill in gaps in precipitation series. Among them, the most widely used are neural network-based methods such as self-organizing maps (Nkiaka *et al.* 2016; Ben Aissia *et al.* 2017; Miró *et al.* 2017; Teegavarapu *et al.* 2017), expectation-maximization algorithms (Schneider 2001; Yozgatligil *et al.* 2013; Ben Aissia *et al.* 2017; Miró *et al.* 2017), multiple imputation by chained equations (Radi *et al.* 2015; Sattari *et al.* 2017; Ben Aissia *et al.* 2017; Burhanuddin *et al.* 2017), copula-based methods (Bárdossy & Pegram 2014; Ben Aissia *et al.* 2017), and spatio-temporal imputation (Teegavarapu 2009; Ben Aissia *et al.* 2017).

In the studies reviewed, the degree of missingness in precipitation time series ranges from low (<1%) to high (50–60%) with an average around 30%. A percentage of missingness above 60% is reported in only one article, for 2 out of the 54 rain gauges used (Teegavarapu *et al.* 2017). However, missingness may often be greater, particularly when longer historical records are considered. One could think of removing those time series from the analysis. However, this is not always an option, as the number of available stations may be too limited or rain gauges with large percentages of MV may be located in areas representative of certain smaller scale hydro-meteorological processes which determine system characterization and modeling.

Precipitation is a semi-continuous variable with a large proportion of days having zero precipitation. Hydrological models are very sensitive to this condition, especially in arid and semi-arid areas where zero rain days are the most frequent. Interpolation methods tend to overestimate the number of rainy days and underestimate extreme events, so that the probability distribution of precipitation is not preserved (Simolo *et al.* 2010; Teegavarapu 2014; Miró *et al.* 2017). However, in many studies, the distributions of precipitation time series are not considered and only performance measures are taken into account to compare imputation methods (Radi *et al.* 2015; Burhanuddin *et al.* 2017). Recently, post-interpolation bias-correction methods based on quantile matching have been used by some authors (Teegavarapu 2014; Miró *et al.* 2017). Nevertheless, these methods cannot be used when the observed sample size is small and the proportion of MV large because there is not sufficient ground to compare the probability distributions of observed and estimated data.

The present study tests three approaches (spatio-temporal kriging, multiple imputation with predictive mean matching and random forest) to filling in missing daily precipitation data in a dataset with 64% of MV and where extreme missingness (>90% MV) is observed in some rain gauges. Different random and chronological missing patterns in the dataset are assessed. To our knowledge, this study is the first to consider extreme missingness (>90%) as well as the RF algorithm to impute precipitation data.

METHODS

Study area and data

A precipitation dataset from 112 rain gauges in the period October 1975–May 2017 (15,219 × 112 matrix) with an overall 64% of missingness was used (Figure 1). The stations are located all over the 2,640 km^2 area covered by the Almonte-Marismas Aquifer in SW Spain, connected to the Doñana National Park wetland system. It is a flat area near the sea where no major orographic barriers are present. Doñana has a sub-humid Mediterranean climate with Atlantic influence. Rainfall is quite variable, with a 580 mm yearly average, about 80% of which is distributed over a wet period from the end of September to the beginning of April. Spatial distribution of rainfall is controlled by Atlantic fronts, which are partially intercepted by small elevations of up to 70 m located near the coast. In addition, the effect of Mediterranean and Atlantic air mass shocks increases precipitation values near the Guadalquivir River.

The distribution of rain gauges according to the percentage of MV is variable and random (Figure 1). Only 25% of the stations have a percentage of MV below 46.7% and 25% of them have more than 89% of MV. Some of the stations with extreme missingness cover wide areas where no other rain gauges are present, such as those located in the central-north and southern parts of the aquifer.

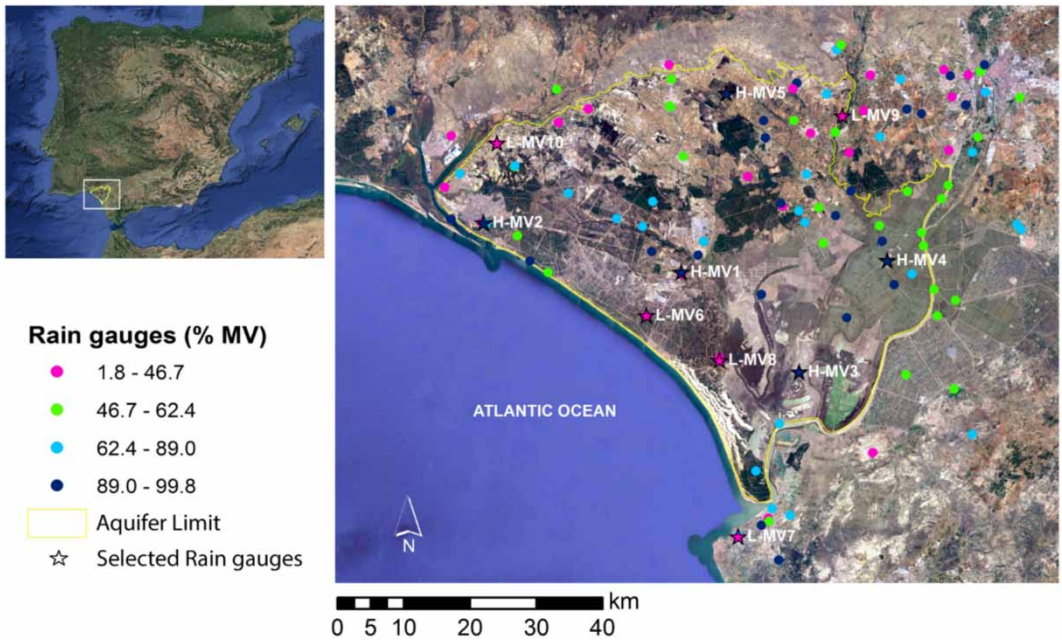

Figure 1 | Location of rain gauges in the study site and boundary of the Almonte-Marismas groundwater system. The number of missing values (MV) in each rain gauge is represented by the quartile interval of the overall distribution of the percentage of missing values. Stars represent the rain gauges selected for model comparison. H-MV, high proportion of missing values; L-MV, low proportion of missing values.

Imputation methods

Three imputation methods (spatio-temporal kriging, multiple imputation with predictive mean matching, and random forest) that can deal with complex non-linear patterns and relations between rain gauges have been selected. All of them have proved to be suited for large-scale imputation in a wide variety of cases, including environmental data (Genton 2007; Radi *et al.* 2015). Furthermore, the three methods are freely available in R programming language (R Core Team 2018).

Spatio-temporal kriging (STK) is a geostatistically based method that takes spatio-temporal correlations into account (Genton 2007). The method estimates a spatio-temporal covariance/variogram model and performs spatio-temporal interpolation (Gräler *et al.* 2016). Among the various types of spatio-temporal covariance structures available, in the present study the separable covariance product model yielded the best results. STK is implemented with the R package *gstat* (Pebesma 2004). *Multiple imputation by chained equations* generates m imputations based on sequential imputation regression models of each variable conditioned by all other variables (van Buuren *et al.* 2006). Applying Rubin's rules, the point estimate is the mean of the m estimates (Rubin 1987). When predictive mean matching (PMM) is used as the estimating regression model, imputed values are sampled only from the k observed values of the respective variable that match predicted values as closely as possible (White *et al.* 2011). Therefore, plausible imputed values are guaranteed. PMM is implemented here with the R package *micemd* (Audigier & Resche-Rigon 2018), which allows for parallel calculations. The last method, *missForest* (Stekhoven & Bühlmann 2012), is a non-parametric iterative imputation method based on the random forest (RF) algorithm (Breiman 2001). It trains a RF on observed values of each variable as a first step, followed by predicting the missing values and then proceeding iteratively until the stopping criterion is met or the user-specified maximum number of iterations is reached. The non-parametric nature of RF has the advantage of not having to make any assumptions about the distributions of data or imputation models. It only requires the observation to be pairwise independent. Although this hypothesis does not hold for daily rainfall, we assume it can be relaxed due to the inherent robustness of the random forest algorithm against correlated variables

by randomly sampling a subset of the variables at each split (Breiman 2001). The *missForest* method is implemented with the R package *missForest* (Stekhoven 2013) and it can be run in parallel to save computation time. For both PMM and RF imputation schemes, the variables are imputed with respect to increasing numbers of MV and predicted values are used in subsequent imputations.

Imputation performance is closely related to the distance between rain gauges. In the case of STK, this relation is explicit, as the spatial correlation has a decreasing trend up to the spatial variogram range. Distance-based correlation is also implicitly accounted for by data-driven methods (PMM and RF) as they search for best predictors across all rain gauges. Nevertheless, in this study, we focus on the amount of MV and missing patterns as key factors for performance that can be directly compared between methods.

Imputation framework

Ideally, a controlled experiment with a complete dataset would provide a useful benchmark to assess the performance of imputation methods. It would also allow comparison of the probability distributions of observed and estimated rainfall under different degrees of missingness. However, in the present study case, when missing values are filtered out, only a small subset of ten rain gauges spread around the limits of the aquifer and 18 years of data remain. This reduced dataset is not representative of the spatial characteristics of the study area. Moreover, a smaller amount of information conditions the performance of data-driven methods (PMM and RF), so results might not be representative of the real situation. In such cases, there is a high risk of not adequately characterizing the performance of methods.

Ten rain gauges located across the study site showing different missingness characteristics were selected to compare the performance of the three methods (Figure 1). The selection criteria were based on spatial coverage of the study area, degree of missingness, missing data patterns, and both presence and absence of nearby rain gauges with fairly complete time series in order to account for spatio-temporal variability. Five of them, named H-MV, have a very high degree of missingness (90% to 98% MV), while the other five (L-MV) show lower amounts of missing information (6% to 25% MV). Summary statistics of these rain gauges along with their percentage of missing values are provided in Table 1 (raw time series and summary statistics from all 112 rain gauges is provided as Supplementary material 1, Table S1). All median values of daily precipitation equal 0. This is related to the positive skewness of the distribution of daily precipitation. Mean and standard deviation are similar in both groups of rain gauges, but the range of observed values is larger in L-MV gauges than in

Table 1 | Summary statistics and percentage of missing values (MV) in daily precipitation series in the period October 1975–May 2017 of the ten rain gauges selected for imputation method comparison

Rain gauge	Mean (mm)	Median (mm)	SD (mm)	Range (mm)	Rainy days (%)	MV (%)	n
H-MV1	1.02	0	5.42	70.20	34.36	97.86	326
H-MV2	1.32	0	6.05	70.50	32.00	97.70	350
H-MV3	2.45	0	7.55	65.00	16.67	96.61	516
H-MV4	1.82	0	6.74	65.00	10.78	95.43	696
H-MV5	2.37	0	7.46	77.00	19.20	90.52	1,443
L-MV6	1.64	0	6.12	101.50	15.40	25.32	11,365
L-MV7	1.50	0	5.91	128.50	19.31	24.21	11,535
L-MV8	1.51	0	5.86	90.00	15.35	13.34	13,189
L-MV9	1.63	0	6.25	112.50	17.06	6.03	14,302
L-MV10	1.53	0	6.15	106.10	13.18	5.66	14,357

SD, standard deviation; n, number of available records.

H-MV. The proportion of rainy days lies between 10% and 20% in most cases, which makes the correct identification of this variable a big challenge.

Two rain gauges, H-MV1 and H-MV2, only have records in 2008 (Figure 2) but there are two stations with the lowest degree of missingness lying very close to each of them (Figure 1). The other three rain gauges with extreme missingness (H-MV3, H-MV4, and H-MV5) are located in the central-north and southern areas with no fairly complete stations nearby (Figure 1). Available data in the more complete stations cover extensive sections of the period studied (Figure 2). Both random (H-MV5, L-MV6, L-MV7, L-MV8, L-MV9, L-MV10) as well as chronological (H-MV1, H-MV2, H-MV3, H-MV4, L-MV6, L-MV8, L-MV10) missing data patterns are observed.

For each of the gauges selected, three sets of train/test splits were carried out on the available data to analyze different missing data patterns (Figure 3). Random missingness was assessed with a 50/50 random partition where 50% of the available data was used as the training set and the remaining 50% of the data as the test set. The time-dependence structure of rainfall variability was accounted for with the other two splits. When the first 50% of the series is kept as observational data for training and the last 50% is estimated and used for testing, the pattern is called 'Last half'. If the last 50% of the data series is used for training and the first 50% of the series is filled for testing, the pattern is referred to as 'First half'. Each imputation method was applied to impute the MV using the training sets for the ten stations and all available data for the remaining 102 stations.

Instead of the usual 80/20 or 75/25 data partition between the train and test sets in predictive models, an ambitious 50/50 split was chosen. The idea behind this decision was to increase the amount of validation data in rain gauges with little available information as well as pushing the capacities of the three methods to the limit.

STK and PMM are capable of imputing zero rainfall values but RF imputes averaged non-zero predictions. However, in multiple imputation with PMM, the average of m imputations is then taken as the estimate of the final value, therefore non-rain days will be missed unless all imputed values equal zero. To adequately fill in zero rain days with the RF and PMM algorithms, a correction with

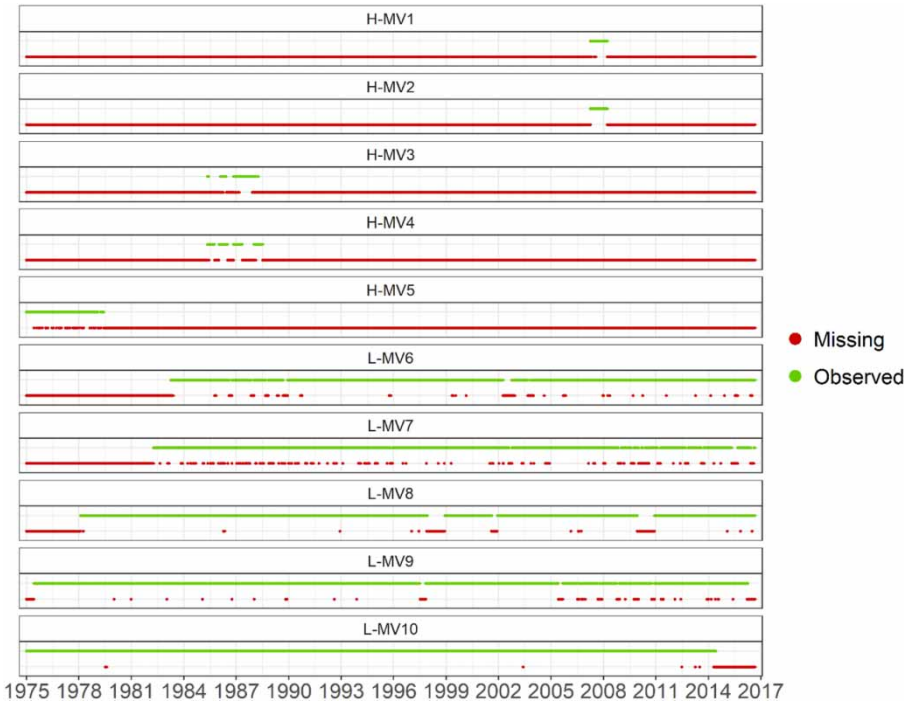

Figure 2 | Available and missing data in the time series of the ten rain gauges used for model evaluation.

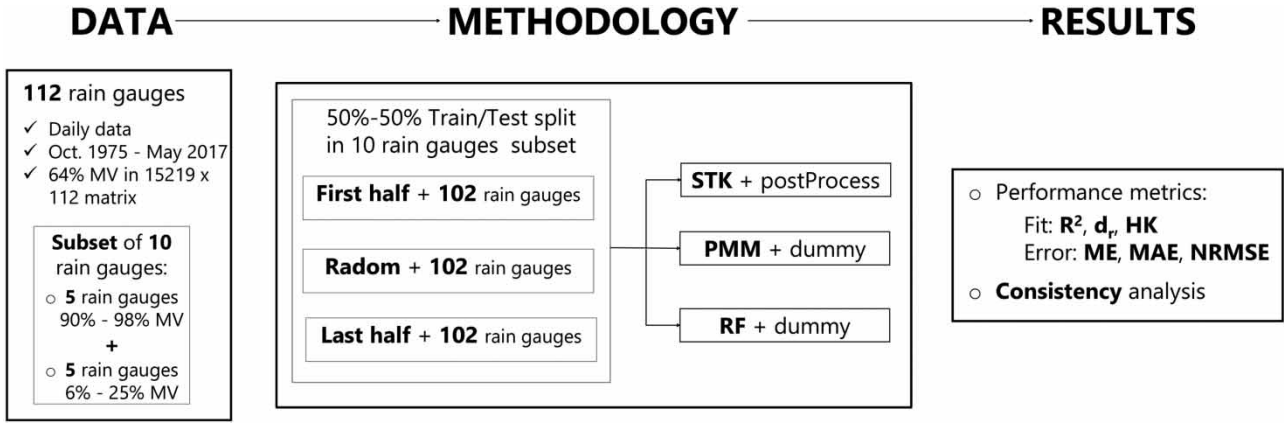

Figure 3 | Methodological scheme of the validation approach for precipitation data imputation. First ten rain gauges with different degrees of missingness are selected for model evaluation. Three train/test splits accounting for different missing data patterns are carried out. Imputation of missing values in all rain gauges is then performed with each method using the training data of the selected rain gauges and all available data of the remaining 102 stations. The results are then compared in terms of several performance metrics. MV, missing values; STK, spatio-temporal kriging; PMM, multiple imputation with predictive mean matching; RF, random forest; R^2, coefficient of determination; d_r, refined index of agreement; HK, Hanssen–Kuipers discriminant; ME, mean error; MAE, mean absolute error; NRMSE, normalized root mean squared error.

dummy variables was introduced as a pre-processing bias–correction strategy. The incomplete data matrix was extended with one dummy variable per rain gauge ($15{,}219 \times 224$) accounting for the presence (1) or absence (0) of rain on a certain day. As both methods can impute categorical variables, first the estimate of the dummy variable for each date was considered and for each rain gauge all dates with 0 value were imputed with 0 rain whereas only for those dates where the dummy variable predicted a value of 1 was the numeric rain estimation imputed.

Based on the literature reviewed, in the PMM scheme, due to the large degree of missingness in the dataset, values of $m = 30$ imputations and $k = 5$ donor pool were used (White *et al.* 2011; Morris *et al.* 2014). These values provided a suitable tradeoff between performance and computational cost. To compute the 30 imputations, the mode was used to get an estimate of the dummy variable on each date and the mean for the quantitative rainfall values, as mentioned above. In terms of computational cost, the inclusion of extra variables in the PMM and RF methods is offset by the parallel processing implemented in the R packages *micemd* and *missForest*.

The spatio-temporal experimental variogram was modeled with a spatial component using a spherical variogram with a 45 km range, and a temporal component using an exponential variogram with a range of 6 days. These ranges indicate that rainfall presents spatial correlation up to 45 km and temporal correlation up to 6 days. The total fitted spatio-temporal sill was of 36 mm^2/day. Measurement errors were taken into account in the standardized spatial and temporal models by means of the partial nuggets (i.e., 0.2 mm^2/day and 0.5 mm^2/day, respectively). The spatio-temporal variogram was then input into the spatio-temporal ordinary kriging algorithm to estimate MV in each rain gauge. Further information on the characteristics of the spatio-temporal variogram are provided as Supplementary Material 2.

RF and STK have the disadvantage that negative rainfall values can arise. This is usually solved with a post-processing correction by assigning zero values to all the negative imputation results.

Evaluation of methods

The performance of the estimation methods used was compared and assessed using the coefficient of determination (R^2), refined index of agreement (d_r), Hanssen–Kuipers discriminant (HK), mean error (ME), mean absolute error (MAE), and normalized root mean squared error expressed as a percentage (NRMSE). The first three are measures of goodness-of-fit and model performance, while the last three are error metrics to measure bias and accuracy.

The dimensionless d_r index is highly consistent compared to other popular indices such as the Nash and Sutcliffe index and suitable for comparison of competing methods, particularly for daily precipitation estimation in arid locations (Willmott *et al.* 2012). It is bounded by −1

and 1 and expressed as:

$$d_r = \begin{cases} 1 - \dfrac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{2\sum_{i=1}^{n} |y_i - \bar{y}|}, & \text{when } \sum_{i=1}^{n} |\hat{y}_i - y_i| \le 2\sum_{i=1}^{n} |y_i - \bar{y}| \\ \dfrac{2\sum_{i=1}^{n} |y_i - \bar{y}|}{\sum_{i=1}^{n} |\hat{y}_i - y_i|} - 1, & \text{when } \sum_{i=1}^{n} |\hat{y}_i - y_i| > 2\sum_{i=1}^{n} |y_i - \bar{y}| \end{cases} \quad (1)$$

where y_i is the observed value of daily rainfall, $\hat{y}_i$ is the imputed value of daily rainfall missing observation, $\bar{y}$ is the mean of the observed values, and n is the number of observations. Values close to 1 indicate good model performance. An advantage of d_r is that it approaches 1 slowly, so it provides greater separation when comparing methods that perform relatively well.

The HK score is used to distinguish between occurrences and non-occurrences of a rain event (Hanssen & Kuipers 1965). The score has a range of −1 to +1, where 0 represents no skill or a random estimate and 1 represents a perfect estimate. Woodcock (1976) argued that the HK is universally acceptable for evaluating yes/no meteorological forecasts. HK is defined as:

$$\mathrm{HK} = \frac{AD - BC}{(A + C)(B + D)} \quad (2)$$

where A, B, C, D are the number of classified rain events as defined in contingency Table 2. HK is widely used in precipitation studies (Teegavarapu 2014; Kim & Ryu 2016).

Error metrics allow for the comparison of the average absolute (MAE) and relative (NRMSE) differences between the observed and the imputed MV. Furthermore, ME provides a measure of bias. They are calculated as:

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i - y_i \quad (3)$$

Table 2 | Success and failure combinations when predicting rain and no rain records in the rain gauges

	Rain observation	No rain observation
Rain estimation	A	B
No rain estimation	C	D

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \quad (4)$$

$$\mathrm{NRMSE} = \frac{\sqrt{\dfrac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}}{y_{max} - y_{min}} * 100 \quad (5)$$

where y_{max} is the maximum observed value and y_{min} is the minimum observed value.

Uncertainty in the estimates of imputed values is generally not treated in depth in hydrological applications (Ben Aissia *et al.* 2017). Uncertainty on those estimated values should be considered for any subsequent application. Methods like copulas provide evaluation of the uncertainty of the estimations through the conditional distribution of precipitation at a selected point (Bárdossy & Pegram 2014). However, it is very difficult to obtain these distributions, especially when most data are missing. Another way to assess the performance of imputation methods is through the uncertainty analysis of calibrated and validated hydrological simulation models. Chen *et al.* (2019) evaluated the impacts of rainfall imputation methods and missing patterns on the uncertainty of flow and total phosphorus model simulations. Nevertheless, if it is not possible to carry out a thorough uncertainty analysis, it is at least necessary to go beyond basic performance measures and check how consistent the imputed precipitation data are. The large degree of missingness of the dataset hampers the comparison of meaningful statistics between the observed and imputed series. Therefore, a qualitative consistency analysis of the results was performed in terms of the monthly and yearly distributions of the imputed precipitation series across methods and missing patterns.

RESULTS AND DISCUSSION

Performance and error analyses

First, results for the test sets are analyzed for each imputation method by percentage of MV (Figure 4). PMM is the most sensitive method to the degree of missingness, showing the worst performance in the presence of large amounts of MV. The highest median MAE and NRMSE (MAE = 1.65 mm,

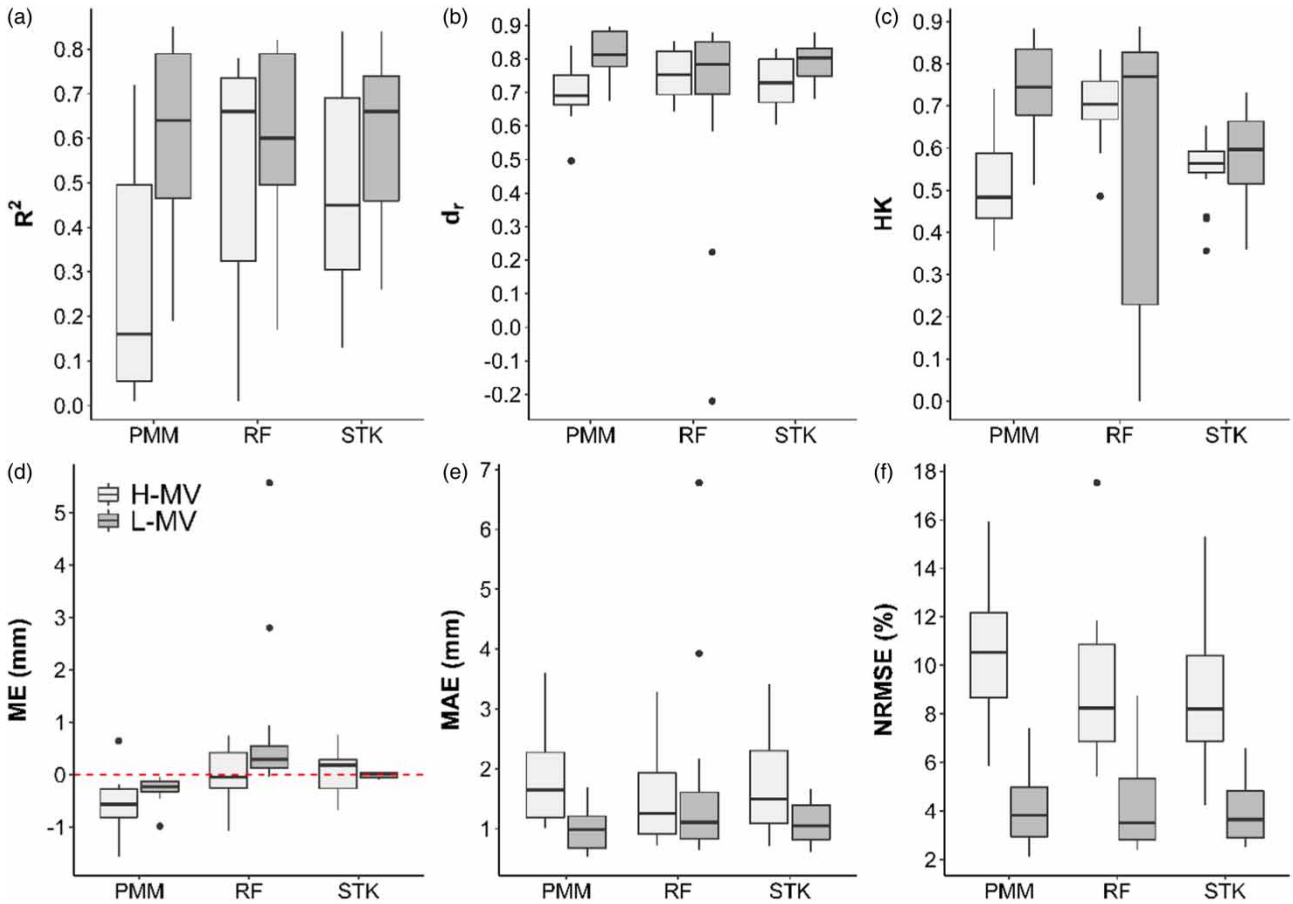

Figure 4 | Distribution of performance measures (test set) of missing precipitation data imputation in ten rain gauges grouped by proportion of missing values and imputation method. MV, missing values; PMM, multiple imputation with predictive mean matching; RF, random forest; STK, spatio-temporal kriging; H-MV, high proportion (between 90% and 98%) of MV; L-MV, low proportion (between 6% and 25%) of MV; R^2, coefficient of determination; d_r, refined index of agreement; HK, Hanssen–Kuipers discriminant; ME, mean error; MAE, mean absolute error; NRMSE, normalized root mean squared error.

NRMSE = 10.5%) and the lowest median R^2 and HK (R^2 = 0.16, HK = 0.48) in H-MV cases are observed for the PMM method (Figure 4(a), 4(c), 4(e) and 4(f)). This is related to the limitations found for PMM with regard to small dataset sizes and large amounts of MV due to the hot-deck characteristics of the method (White *et al.* 2011; Morris *et al.* 2014). STK and RF show similar accuracy in the H-MV imputations in terms of NRMSE (median values around 8%), but the latter shows lower MAE (median values are 1.49 mm and 1.26 mm for STK and RF, respectively), and better performance measures (median R^2 is 0.45 for STK and 0.66 for RF; median HK is 0.56 for STK and 0.70 for RF).

The bias analysis in terms of ME shows three patterns (Figure 4(d)): underestimation in PMM in both groups of imputations (median ME is −0.56 mm for H-MV and −0.23 mm for L-MV), overestimation in RF in cases with fewer MV (median ME is −0.05 mm for H-MV and 0.30 mm for L-MV), and slight overestimation in STK in the presence of large missingness (median ME is 0.18 mm for H-MV and 0.01 mm for L-MV).

The d_r index of agreement indicates overall good model performance (Figure 4(b)). The lowest median d_r is 0.69 for PMM in H-MV imputations and the highest median d_r is 0.81 for PMM in L-MV, d_r values of RF and STK lying between these two. The index indicates that RF is less sensitive to the amount of MV, as similar distributions are observed for H-MV and L-MV, whereas STK and PMM show increased d_r values in the L-MV group.

L-MV imputations show considerable improvements in most metrics compared to H-MV imputations (median $R^2 \geq 0.60$, median HK $\geq$ 0.60, median MAE $\leq$ 1.11 mm,

and median NRMSE $\leq$ 4%), particularly for PMM. Caution should be exercised when assessing the large differences in NRMSE between H-MV and L-MV imputations. Normalization in NRMSE is carried out with the range of data observed (Equation (5)), meaning that rain gauges with small amounts of observed data may have the denominator reduced and, thus, increased NRMSE (note range values in Table 1).

There were two rain gauges in the L-MV case where RF underperformed, L-MV7 and L-MV8, entailing a wider distribution of HK and the presence of outliers in the distributions of d_r, ME, and MAE (Figure 4(b)–4(e)).

Figure 5 represents the distribution of performance measures in the test sets for each imputation method by type of missing data pattern structure. It can be seen that the poor performance of RF is mainly related to the 'First half' partitions, where the first 50% of the available data of each series were removed before imputation and then used for model testing. Positive bias in this missing pattern (median ME = 0.41 mm) doubles that of the 'Last half' (median ME = 0.20 mm) and is almost four times higher than ME in the random partition (median ME = 0.12 mm). Furthermore, HK and MAE show wider distributions towards lower and higher values, respectively (Figure 5(c) and 5(e)). The underlying reason is the larger joint missingness in the dataset in the first part of the period which limits the learning capability of the RF algorithm (i.e., data sparsity).

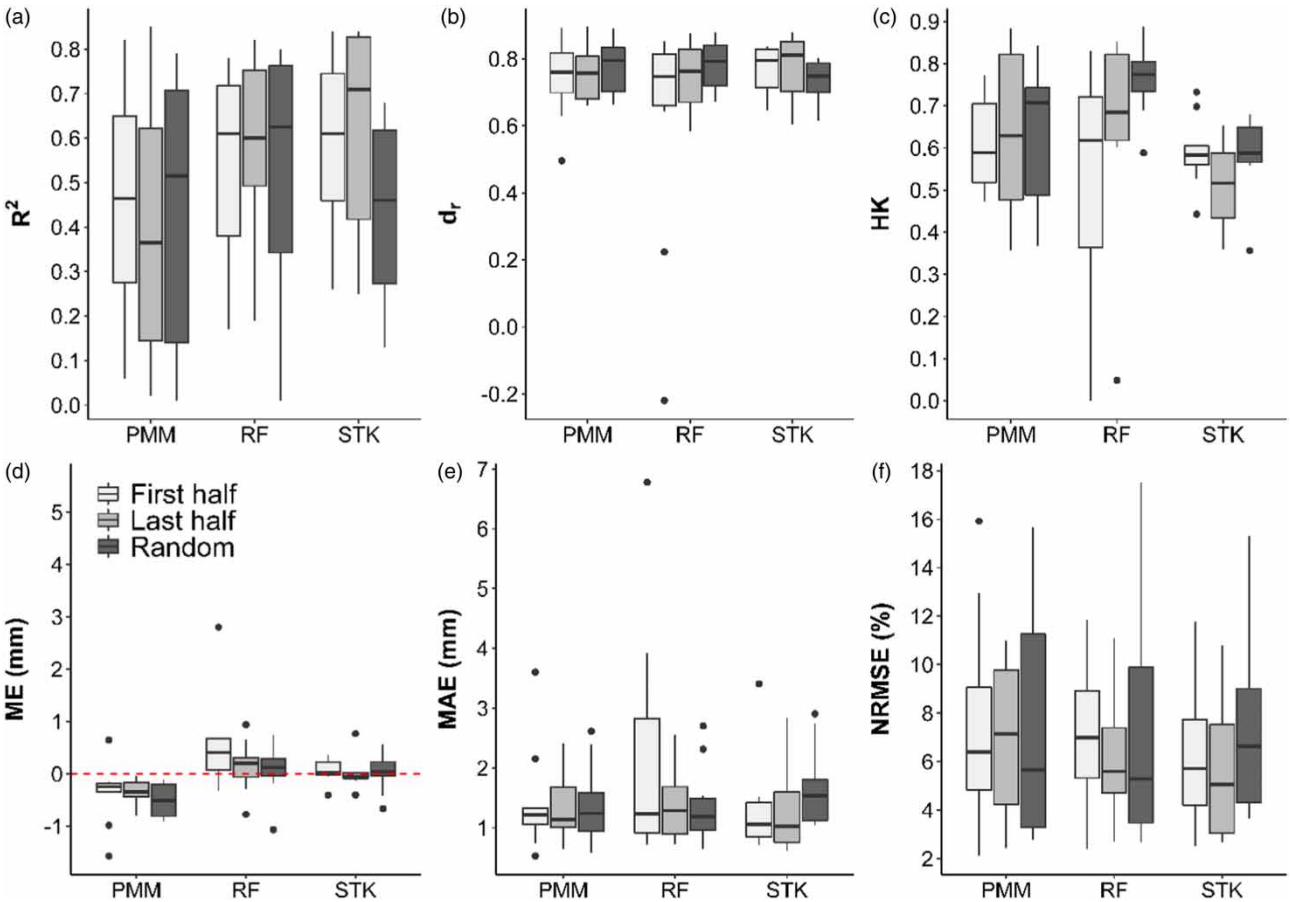

Figure 5 | Distribution of performance measures (test set) of missing precipitation data imputation in ten rain gauges grouped by missing pattern (train/test data splits) and imputation method. PMM, multiple imputation with predictive mean matching; RF, random forest; STK, spatio-temporal kriging; R^2, coefficient of determination; d_r, refined index of agreement; HK, Hanssen–Kuipers discriminant; ME, mean error; MAE, mean absolute error; NRMSE, normalized root mean squared error. 'Random' refers to 50/50 random data partition in train and test datasets for each rain gauge. When the first 50% of available data in each series is kept as training data and the last 50% is estimated and used for testing, the pattern is called 'Last half'. If the last 50% of the data is kept as training data and the first 50% is estimated and used for testing, the pattern is referred to as 'First half'.

The previously observed bias towards underestimation in PMM is stronger in the case of random missing patterns (Figure 5(d)). Median ME values across data partitions in PMM are −0.51 mm for 'Random', −0.34 mm for 'Last half' and −0.24 mm for 'First half'. This negative bias is the effect of averaging within the range of observations in the multiple imputation scheme. It also yields relatively constant absolute error (MAE) and d_r across missing patterns (Figure 5(b) and 5(e)). However, the poorer performance of PMM compared to STK and RF is evidenced by the wider distributions of R^2 and NRMSE for all missing patterns (Figure 5(a) and 5(f)).

Again, the HK plot in Figure 5(c) corroborates the successful introduction of dummy variables to account for the occurrence of rainy and non-rainy days in data-driven methods compared to the spatio-temporal based STK method. Median values of the HK score across missing patterns range from 0.52 to 0.59 for STK and from 0.62 to 0.77 in RF.

The lowest performance of STK (median $R^2 = 0.46$, median MAE = 1.54 mm, median NRMSE = 6.6%, median $d_r = 0.75$) is observed in imputations of random splits, where RF has a very good performance (median $R^2 = 0.63$, median HK = 0.77, median MAE = 1.18 mm, median NRMSE = 5.3%, median $d_r = 0.79$). Conversely, chronologically ordered imputations are better accomplished through STK.

The correction with dummy variables in the RF and PMM algorithms implied doubling the number of variables. However, the possibility of carrying out parallel calculations with the R packages *micemd* and *missForest* saved computing time. Total processing times were 7 hours for PMM, 12 hours for RF, and 10 hours for STK on a system with Intel(R) Core(TM) i7- 4790 CPU @ 3.60 GHz, x64-based processor and 7.98 GB usable RAM. STK has large computational cost due to the computational complexity of matrix inversion.

Table 3 shows the best imputation method for each rain gauge and missing pattern according to the performance and error measures analyzed. The criterion for method selection was based on ranked goodness-of-fit measures ($HK > R^2 > d_r$) and on ranked error measures (MAE > NRMSE > ME). The order of the measures was decided upon their discriminatory capacity shown in Figures 4 and 5, and on their relative importance for the phenomenon (i.e., a higher weight was given to HK due to the importance of correctly assigning rainy/non-rainy days). Out of the 30 models, RF was the best option in 16 cases, STK in 9 cases, and PMM in 5 cases. RF is the preferred method in the presence of large amounts of MV (H-MV rain gauges) and in cases of random missing patterns. STK is selected in time-dependent missing patterns in both H-MV and L-MV gauges. PMM is only suitable for imputing rainfall data in cases where the percentage of MV is below 25%.

In summary, the results are encouraging given the large degree of missingness and they support the suitability of the introduction of dummy variables to impute non-rainy days. Performance and error measures are comparable and often better than those achieved in other studies infilling daily precipitation data using different techniques (Lo Presti *et al.* 2010; Simolo *et al.* 2010; Bárdossy & Pegram 2014; Teegavarapu 2014; Radi *et al.* 2015; Kim & Ryu 2016; Burhanuddin *et al.* 2017; Teegavarapu *et al.* 2017; Jahan *et al.* 2018). All these studies consider much lower degrees of missingness, but results are highly dependent on the network structure and the climatological conditions.

Table 3 | Best performing method in missing precipitation data imputation for each rain gauge and missing pattern according to performance and error measures

Gauge	H-MV1	H-MV2	H-MV3	H-MV4	H-MV5	L-MV6	L-MV7	L-MV8	L-MV9	L-MV10
Pattern										
First half	RF/STK	STK	RF	STK	RF	STK	PMM	STK	PMM	RF
Last half	RF/STK	STK	RF	RF	RF/STK	STK	STK	STK	PMM	PMM
Random	STK	RF	RF	RF	RF	RF/STK	PMM	RF	RF	RF

Acronyms to the right of the slash symbol represent final selected methods after consistency analysis (section below).
H-MV, high percentage of missing values; L-MV, low percentage of missing values; PMM, multiple imputation with predictive mean matching; RF, random forest; STK, spatio-temporal kriging.

Both the volume of MV and the missing patterns condition the performance of the imputation methods. PMM is not suitable for imputing precipitation datasets with extreme missingness and limited observations. Both RF and STK show similar performance regarding the degree of missingness. However, the spatio-temporal structure of STK provides better results with time-dependent missing patterns, whereas RF is the best method to impute random missingness. STK is spatially based and the spatio-temporal separable covariance function accounts for intrinsic processes in the data structure that hold when sequential data are available. That is the reason why it performs better under chronological missing data patterns than under random missing data patterns. Conversely, the opposite holds for the data-driven RF method.

Finally, PMM has a generalized negative bias (i.e., underestimation) due to the imputation of averaged observed values, while RF tends to overestimate observations under chronological missing data patterns. On a case-by-case basis, RF was the best method in 50% of the tests carried out (Table 3).

Consistency analysis

In real-world situations, we should expect a combination of random and chronological missing data patterns in precipitation datasets (Figure 2). From the results in the previous section, the RF algorithm emerges as a valid imputation method in cases of extreme missingness provided that sufficient information from other rain gauges is available. Based on the HK score, the inclusion of dummy variables allows for successful determination of the occurrence of rainy and non-rainy days (Figures 4(c) and 5(c)). However, the RF method showed bias towards overestimation under chronological missing patterns (Figure 5(d)). This could pose a serious risk for hydrological modeling if bias is systematically propagated through the time series. However, error measures depend on the particular train/test split and systematic bias is not guaranteed.

To analyze the consistency of the imputations we have aggregated the complete dataset (1975–2017) of the imputed series by hydrologic year and by month (Figure 6 and Figure S1 in Supplementary material 3). Annual precipitation is the sum of daily precipitation for each year and monthly is the average of monthly totals. Consistency analysis was based on the visual inspection of the plots in Figure 6 allowing detection of systematic bias in the imputed series. Large overestimations can be observed with the RF method in some cases of chronological missing patterns. Specifically, based on performance and error measures, RF was the selected imputation method in H-MV1 for the 'First half' and 'Last half' missing patterns and in H-MV5 for the 'Last half' partition (Table 3). However, annual and monthly aggregates show the inadequacy of the method in these cases (Figure 6(a)–6(d)). The underperformance of RF in the 'First half' estimation of H-MV1 results in overestimation of annual precipitation in the first part of the period. This has a particular effect on the summer months (June, July, and August), for which monthly averages exceed 20 mm (Figure 6(a)) when values close to zero are expected. In the 'Last half' case of H-MV1, we find the opposite circumstance; the RF method appears to underestimate precipitation from the 1990s onwards, taking into account that the average annual precipitation in the Doñana area is higher than 500 mm and that the ME for this train/test split was −0.77 mm.

The overestimation problem is worse for H-MV5. Observations in this rain gauge were recorded in the period 1975–1980 (Figure 2). During this period all methods yield similar annual estimations, but from the mid-1980s onwards, RF under the 'Last half' missing pattern severely overestimates precipitation (Figure 6(c) and 6(d)).

The RF method overestimated precipitation data in the 'First half' missing pattern of L-MV7 and L-MV8 (Figure 6(g)–6(j)), as was detected by performance and error measures. In the L-MV group, RF was the best choice under random missing patterns (Table 3). In these cases, monthly and annual imputed precipitation series show consistent results, except for some overestimation in L-MV6 during the first years of the period analyzed (RF_R lines in Figure 6(f)).

The RF algorithm outperformed other imputation methods with mixed data-type datasets (Stekhoven & Bühlmann 2012). Tang & Ishwaran (2017) tested different RF missing data algorithms on 60 datasets and found good robust performance under moderate to high missingness. They also observed that performance improved with increasing correlation. This might be a partial cause for the bad

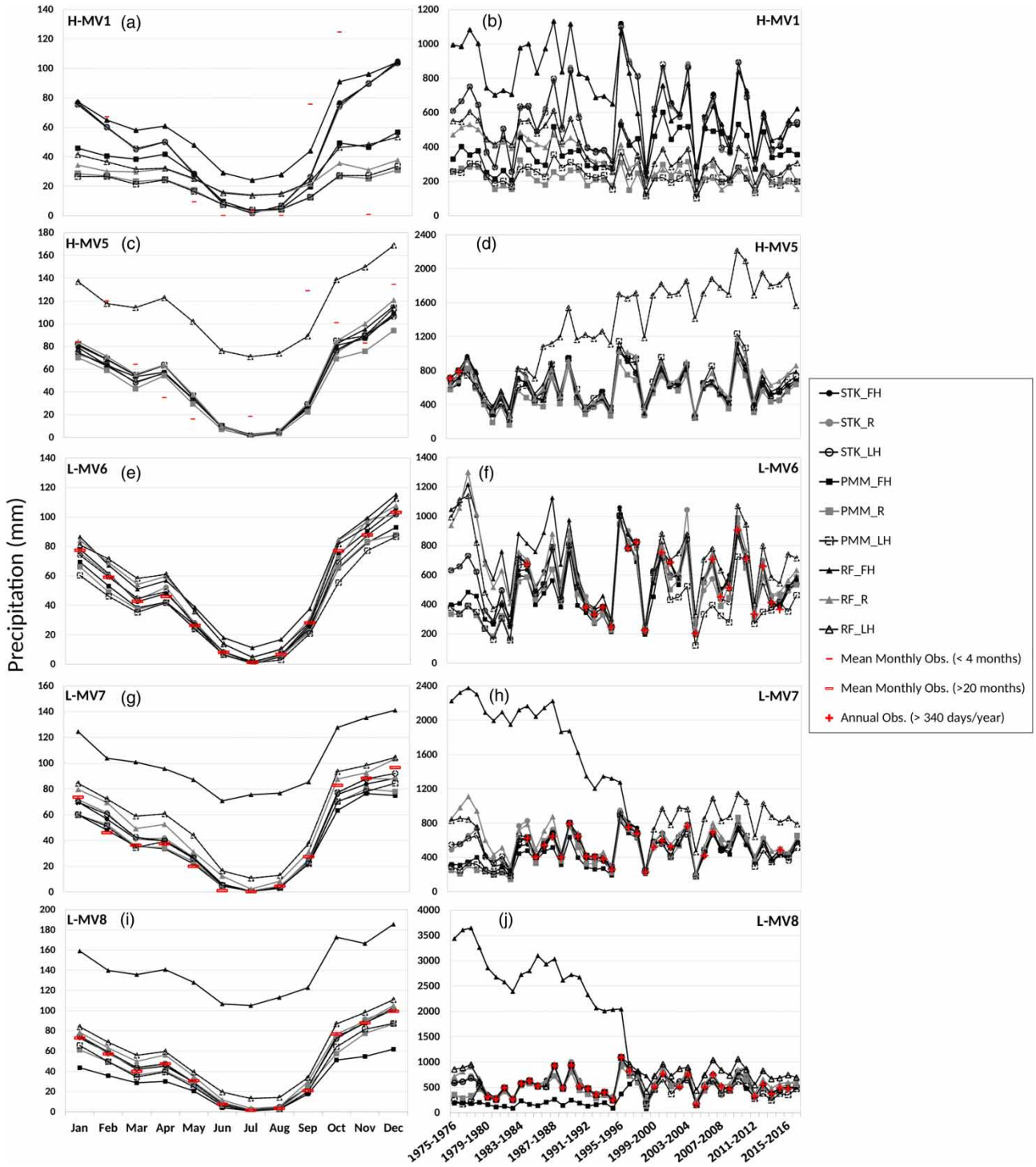

Figure 6 | Mean monthly (a), (c), (e), (g), (i) and total annual (b), (d), (f), (h), (j) precipitation in imputed series of rain gauges with high proportion (H-MV) and low proportion (L-MV) of missing values. The small hyphen symbols account for mean monthly values in H-MV rain gauges having up to three months with records in the time series. The larger rectangle-shaped symbols account for mean monthly values in L-MV rain gauges having a minimum of 20 months with records in the time series. The plus symbols represent observed total annual precipitation in L-MV gauges that have a minimum of 340 observations per year. The plots include all imputation methods and missing data patterns analyzed. Note the last value of the annual precipitation plots only considers the period from October 2016 to May 2017. STK, spatio-temporal kriging; PMM, multiple imputation with predictive mean matching; RF, random forest; FH, first half; R, random; LH, last half.

performance of RF in the cases we are considering. Chronological missing patterns limit the ability of the data-driven model to learn the correlation structure in daily precipitation time series. The dummy variable estimation procedure becomes very sensitive to this issue with the result that non-rainy days go undetected and instead a predicted amount is imputed. Another important result drawn from the H-MV plots in Figure 6 is that RF is highly sensitive to the available information. The few months with observed data in H-MV1 and H-MV5 represent wet autumn periods. When information in other rain gauges is also missing under chronological missing patterns, the RF algorithm is forced to 'learn' these exceptional wet patterns as the 'typical' situation and extrapolate them to other periods (see also Figure S1 in Supplementary material 3).

Underestimation with PMM in the presence of large proportions of MV is evident for all missing pattern scenarios. Its performance improves as the proportion of MV decreases but monthly and annual values remain in the lower ranges (Figure 6).

The STK method shows the most consistent results across missing patterns and proportions of MV (Figure 6). No apparent bias towards under- or overestimation is observed. Slight differences between imputed series with different missing patterns are only present in L-MV6 and L-MV7 (Figure 6(e)–6(h)). The separable covariance product model is, thus, a robust spatio-temporal covariance structure for missing precipitation interpolation under extreme missingness independently of the missing pattern. Although some rain gauges are distant from all the others, they are still within the range of the spatial variogram (i.e., 45 km), which quantifies the extent of the spatial correlation. Therefore, in this site, STK can provide consistent imputations even if the distance to the target rain gauge is large, but always within the variogram range (see, for example, L-MV6 in Figures 1, 6(e) and 6(f)). Nevertheless, the HK score validation results suggest that STK is less accurate in identifying rainy and non-rainy days (Figures 4(c) and 5(c)). This is not problematic for water management on monthly to multi-annual time scales as very small rainfall amounts are imputed in false positive non-rainy day cases. However, it must be taken into account for finer time-scale analyses (i.e., fast conduit flow in karstic systems, surface water–groundwater interactions, etc.). Teegavarapu (2009) showed that a combination of association rule mining with spatial interpolation methods such as ordinary kriging reduced the overestimation of rainy days. However, in that study, only 15 rain gauges were used and the number of association rules grows exponentially based on the number of stations. Further improvements on STK for missing precipitation estimation could be achieved with the development of spatio-temporal indicator kriging implementations.

Given the information provided by the consistency analysis, some corrections were made in the method choices (Table 3). After the update, STK overtakes RF as the best method for estimating missing daily precipitation data in southwest Spain.

Taking into account the computational costs of each method, our results support using STK for missing precipitation data imputation under extreme missingness and time-dependent missing patterns. RF is a suitable choice under any degree of missingness with random missing patterns. Finally, PMM should only be used in cases of lower amounts of MV, but some risk of underestimation still exists due to the effect of averaging multiple imputations.

CONCLUDING REMARKS

The reconstruction of daily precipitation time series for hydrological modeling is a delicate task. In this study, we have tackled the added difficulty of estimating extremely large amounts of MV (overall 64% MV), which is a challenging pre-processing step for fine scale spatio-temporal analysis. Besides suitable performance and error measures, additional considerations must be taken into account to fully describe the phenomena under study. In the results presented, the ME suggested a slight overestimation problem with the RF method, particularly in the 'First half' simulations (Figure 5(d)), but this fact could be masked by other metrics and/or not be systematic. However, when the whole imputed dataset is aggregated in a monthly and annual basis, it can be seen that overestimation is propagated throughout the series, entailing potential mistakes in surface water and groundwater modeling. Therefore, when uncertainty analysis is not feasible, at least consistency of the imputations must be checked before validating the

results of a given method. We found this is not a common practice in some precipitation imputation studies and, thus, we strongly recommend including some kind of consistency analysis for method validation.

In our study, we were able to use a big network of 112 rain gauges covering an area of around 3,000 km^2. This will not always be the case, especially in regions where weather monitoring networks are scarce and sparse. In such circumstances, distant rain gauges need to be considered to test whether the imputation method is capable of finding regional correlations and non-linear patterns at larger regional scales.

STK simulates the distribution of precipitation under chronological missing patterns more consistently than RF and PMM, at the expense of higher computing times. Overall, encouraging results were obtained through the application of these techniques to extreme missingness. With the available data, STK provided consistent results for two rain gauges that only had data in one year in a series of 42 (98% MV). In fact, we should expect improved results of the imputation methods if all available data were used, as 50% of the data in ten rain gauges was removed for model comparison.

The *missForest* RF algorithm emerges as a computationally efficient alternative method for daily missing precipitation estimation when the missing pattern is predominantly random. A correction with dummy variables allowed non-physically based algorithms to determine zero rain days, a crucial factor for hydrological modeling. Further developments in data-driven methods such as taking seasonality in precipitation into account could provide more reliable estimations.

ACKNOWLEDGEMENTS

This research has been funded by CLIGRO Project (MICINN, CGL2016-77473-C3-1-R) of the Spanish National Plan for Scientific and Technical Research and Innovation. This research is also part of the activities subsidized within the National System of Youth Guarantee (PEJ-2014-85121), with financial resources from the Youth Employment Initiative (YEI) and the European Social Fund (ESF); and by the Ministry of Education, Youth and Sport of the Community of Madrid with (IND2017/AMB-7789) for Industrial PhDs. The authors thank the Spanish Meteorological Agency, the Biological Station of Doñana and Junta de Andalucía for the data provided for this work. We would also like to thank the reviewers, whose comments have substantially improved the quality of the paper. We give special thanks to Nuria Naranjo-Fernández for her graphical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this paper is available online at https://dx.doi.org/10.2166/hydro.2020.127.

REFERENCES

Audigier, V. & Resche-Rigon, M. 2018 micemd: Multiple Imputation by Chained Equations with Multilevel Data. R package version 1.2.0. https://CRAN.R-project.org/package=micemd

Bárdossy, A. & Pegram, G. 2014 Infilling missing precipitation records – A comparison of a new copula-based method with other techniques. *Journal of Hydrology* **519**, 1162–1170.

Ben Aissia, M. A., Chebana, F. & Ouarda, T. B. M. J. 2017 Multivariate missing data in hydrology – review and applications. *Advances in Water Resources* **110**, 299–309. https://doi.org/10.1016/j.advwatres.2017.10.002.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32. https://doi.org/10.1023/A:1010933404324.

Burhanuddin, S. N. Z. A., Deni, S. M. & Ramli, N. M. 2017 Normal ratio in multiple imputation based on bootstrapped sample for rainfall data with missingness. *International Journal of GEOMATE* **13** (36), 131–137. http://dx.doi.org/10.21660/2017.36.2760.

Chen, L., Xu, J., Wang, G. & Shen, Z. 2019 Comparison of the multiple imputation approaches for imputing rainfall data series and their applications to watershed models. *Journal of Hydrology* **572**, 449–460.

Genton, M. C. 2007 Separable approximations of space-time covariance matrices. *Environmetrics* **18**, 681–695. https://doi.org/10.1002/env.854.

Gräler, B., Pebesma, E. & Heuvelink, G. 2016 Spatio-temporal interpolation using gstat. *The R Journal* **8** (1), 204–218.

Hanssen, A. W. & Kuipers, W. J. A. 1965 *On the relationship between the frequency of rain and various meteorological parameters*, Report K.N.M.I. 102-81, Communications and Discourses, Royal Netherlands Meteorological Institute, The Netherlands.

Harel, O. & Zhou, X. H. 2007 Multiple imputation: review of theory, implementation and software. *Statistics in Medicine* **26**, 3057–3077. https://doi.org/10.1002/sim.2787.

Jahan, F., Sinha, N. C., Mahfuzur Rahman, M. d., Morshadur Rahman, M. d., Sanaul Haque Mondal, M. d. & Ataharul Islam, M. 2018 Comparison of missing value estimation techniques in rainfall data of Bangladesh. *Theoretical and Applied Climatology* **136** (3–4), 1115–1131. https://doi.org/10.1007/s00704-018-2537-y.

Kim, J. W. & Ryu, J. H. 2016 A heuristic gap filling method for daily precipitation series. *Water Resources Management* **30** (7), 2275–2294. https://doi.org/10.1007/s11269-016-1284-z.

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**, 1–22. https://doi.org/10.1007/s10661-008-0653-3.

Miró, J. J., Caselles, V. & Estrela, M. J. 2017 Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric Research* **197**, 313–330. https://doi.org/10.1016/j.atmosres.2017.07.016.

Morris, T. P., White, I. R. & Royston, P. 2014 Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology* **14**, 75. https://doi.org/10.1186/1471-2288-14-75.

Nkiaka, E., Nawaz, N. R. & Lovett, J. C. 2016 Using Self-Organizing Maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. *Environmental Monitoring and Assessment* **188** (7), 400. https://doi.org/10.1007/s10661-016-5385-1.

Pebesma, E. J. 2004 Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* **30**, 683–691.

Radi, N. F. A., Zakaria, R. & Azman, M. A. 2015 Estimation of missing rainfall data using spatial interpolation and imputation methods. *AIP Conference Proceedings* **1643**, 42–48. https://doi.org/10.1063/1.4907423.

R Core Team 2018 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (accessed 28 June 2019).

Rubin, D. B. 1987 *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, USA.

Sattari, M. T., Rezazadeh-Joudi, A. & Kusiak, A. 2017 Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research* **48** (4), 1032–1044. https://doi.org/10.2166/nh.2016.364.

Schneider, T. 2001 Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* **14**, 853–871. https://doi.org/10.1175/1520-0442(2001)014%3C0853:AOICDE%3E2.0.CO;2.

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**, 1564–1576. https://doi.org/10.1002/joc.1992.

Stekhoven, D. J. 2013 missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4. https://cran.r-project.org/web/packages/missForest/missForest.pdf

Stekhoven, D. J. & Bühlmann, P. 2012 Missforest – nonparametric missing value imputation for mixed-type data. *Bioinformatics* **28** (1), 112–118. https://doi.org/10.1093/bioinformatics/btr597.

Tang, F. & Ishwaran, H. 2017 Random forest missing data algorithms. *Statistical Analysis and Data Mining* **10** (6), 363–377. https://doi.org/10.1002/sam.11348.

Teegavarapu, R. S. V. 2009 Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. *Journal of Hydroinformatics* **11** (2), 133–146. https://doi.org/10.2166/hydro.2009.009.

Teegavarapu, R. S. V. 2014 Statistical corrections of spatially interpolated missing precipitation data estimates. *Hydrological Processes* **28**, 3789–3808. https://doi.org/10.1002/hyp.9906.

Teegavarapu, R. S. V., Aly, A., Pathak, C. H., Ahlquist, J., Fuelberg, H. & Hood, J. 2017 Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: use of optimal weighting parameters and nearest neighbour-based corrections. *International Journal of Climatology* **38** (2), 776–793. https://doi.org/10.1002/joc.5209.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, K. & Rubin, D. B. 2006 Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76**, 1049–1064.

White, I. R., Royston, P. & Wood, A. M. 2011 Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* **30** (4), 377–399. https://doi.org/10.1002/sim.4067.

Willmott, C. J., Robeson, S. M. & Matsuura, K. 2012 A refined index of model performance. *International Journal of Climatology* **32**, 2088–2094. https://doi.org/10.1002/joc.2419.

Woodcock, F. 1976 The evaluation of yes/no forecasts for scientific and administrative purposes. *Monthly Weather Review* **104** (10), 1209–1214. https://doi.org/10.1175/1520-0493(1976)104%3C1209:TEOYFF%3E2.0.CO;2.

Yozgatligil, C., Aslan, S., Iyigun, C. & Batmaz, I. 2013 Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and Applied Climatology* **112**, 143–167. https://doi.org/10.1007/s00704-012-0723-x.

First received 4 July 2019; accepted in revised form 22 January 2020. Available online 28 February 2020