Scientific Research Publishing

Journal of Sensor Technology, 2025, 15(1), 1-13
https://www.scirp.org/journal/jst
ISSN Online: 2161-1238
ISSN Print: 2161-122X

Multi-Task Gaussian Process for Imputing Missing Daily Rainfall Data Using Nearby Stations: Case of Burkina Faso

Souleymane Zio[1], Dazangwende Emmanuel Poan[2], Yoda Adaman[2], Kima Bénéwendé Yves Arnaud Noe[1]

[1]Ecole Polytechnique de Ouagadougou, Ouagadougou, Burkina Faso
[2]Agence National Meteorologie, Ouagadougou, Burkina Faso
Email: ziosouleymane27@gmail.com

How to cite this paper: Zio, S., Poan, D.E., Adaman, Y. and Noe, K.B.Y.A. (2025) Multi-Task Gaussian Process for Imputing Missing Daily Rainfall Data Using Nearby Stations: Case of Burkina Faso. Journal of Sensor Technology, 15, 1-13.
https://doi.org/10.4236/jst.2025.151001

Received: December 12, 2024
Accepted: January 21, 2025
Published: January 24, 2025

Abstract

Precipitation is a critical meteorological factor that significantly impacts agriculture in the sub-Saharan and Sahelian regions of Africa. Accurate knowledge of precipitation levels aids in planning effective agricultural strategies. However, these regions often face challenges with missing rainfall data at numerous gauges. This issue arises due to various factors, including socio-political instability (e.g., terrorism), economic constraints (e.g., insufficient station density due to limited resources), and human factors such as a shortage of qualified personnel. This study evaluates the effectiveness of the multi-task Gaussian process (MTGP) based on the linear model of coregionalization (LMC) for imputing missing daily rainfall data in Burkina Faso, leveraging the correlations among nearby stations. The proposed method is compared with commonly used statistical and machine learning techniques, including mean imputation (ME), K-nearest neighbors (KNN), Multivariate Imputation by Chained Equations (MICE), and Last Observation Carried Forward (LOCF). The results demonstrate that the MTGP approach outperforms MICE, KNN, LOCF, and ME. Additionally, when compared to the independent Gaussian process (IGP), which does not account for correlations between stations, MTGP shows a performance improvement of 50%.

Keywords

Rainfall, Missing Data, Multi-Task Gaussian Process, Correlation

1. Introduction

Meteorological data collected directly from meteorological stations provide essential

time series information critical for identifying extreme climatic events and planning various human activities such as agriculture, fishing, livestock management, aeronautics, and aerospace operations. However, data collection can be interrupted for various reasons, leading to gaps in the dataset. These missing data points can introduce biases in statistical analyses and hinder the development of tools and models requiring continuous time series data across different temporal scales (e.g., daily, monthly, or yearly). In Burkina Faso, rainfall data is recorded by ten (10) synoptic stations and a network of rain gauge stations distributed across the country. However, many of these rain gauge stations, as well as some synoptic stations, experience significant data gaps, particularly in rainfall measurements.

In this study, we investigated the efficiency of multiple task Gaussian process to fill daily rainfall data and, compared it to the statistics approach and well-known machine learning technique as the K-nearest neighbor (KNN) [1] [2], the last observation carried forward(LOCF) [3] and the multivariate imputation by chained equation (MICE) [4] [5] and mean imputation(ME). In ME, we replace the missing values with the corresponding global mean. The Multivariate Imputation by Chained Equations (MICE) [6] begins by first initializing the missing values arbitrarily and then estimating each missing variable based on the chain equations. MICE has been extensively used to fill the missing data in meteorology area [7] [8]. The k-nearest neighbor (KNN) [9] finds similar samples, and imputes the missing values with a weighted average of its neighbors. The multi-output or multi-task learning framework has been successfully applied to various time series prediction problems [9]-[14]. Herein we used the linear coregionalization model (LMC) [9] to consider the correlation between nearby station. We compared the different LMC models as Intrinsic coregionalization model (ICM) [11] and the best model is used to fill daily rainfall data.

The main contributions of the present work, has been the use of multi-task GP framework to fill the daily rainfall data in BF stations, considering the correlation between nearby stations. The multitask Gaussian results are compared to the well-known statistics and machine learning techniques imputation methods. The main contributions of the present work, has been the use of multi-task GP framework to fill the daily rainfall data in BF stations, considering the correlation between nearby stations.

2. Dataset

In this section, we describe the dataset used for our study. The data consists of in-situ rainfall measurements collected from ground-based meteorological stations in Burkina Faso (BF), provided by the National Meteorological Agency of Burkina Faso (ANAM). ANAM operates various types of stations, including 10 synoptic stations and numerous rainfall gauges spread across the country for agricultural purposes. The 10 synoptic stations monitor various meteorological variables and serve as the primary sources of climatic data in BF. These stations rarely experience

missing data. However, the rainfall gauge sensors often suffer from significant data gaps. As illustrated in Figure 1, the distribution of missing data across 147 meteorological stations, which include both synoptic stations and rainfall gauges, is shown for the period from January 2018 to January 2023. The results indicate that the 10 synoptic stations (marked in red) have no missing data, with a 0% missing rate, while all 137 rainfall gauge stations experience varying levels of missing data.

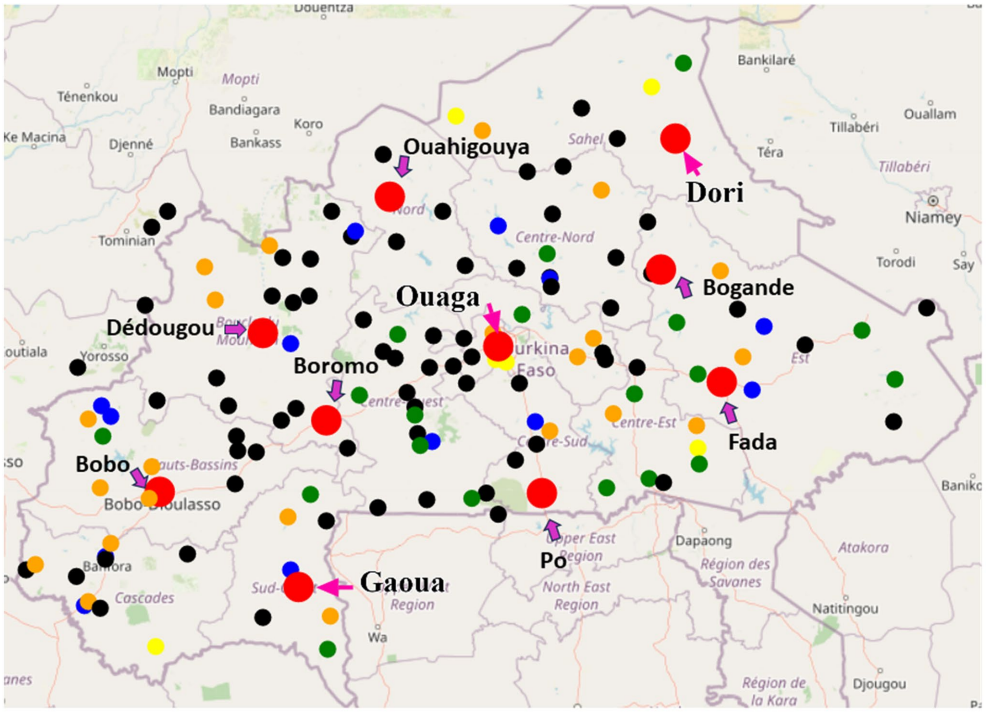

Figure 1. Percentage of rain gauge observation data missing in some meteorology station around Burkina Faso. red colour indicate that 0 % of missing, black colour represented missing between [1 to 10] % of missing. Green colour between [10 to 20] % of missing, orange [20 to 30] % of missing, yellow circle between [30 to 40] % of missing and blue with missing > 40%.

The rainfall data with no missing values need to be available from various stations across the country to build projection models and address key issues such as climate change. However, in addition to the low density of meteorological stations in Burkina Faso (BF), there is a significant amount of missing data in the existing stations. To address this, we propose a method capable of imputing the missing data using information from nearby stations. The proposed methodology relies on a Gaussian process, which takes into account the correlation between neighboring stations to impute the missing daily rainfall data. In the next section, we introduce the Gaussian process.

3. Multi-Output Coregionalization for GPs

Gaussian processes can be used to predict a single-output or multi-output learning problem involving M tasks. In our study, a single output can be represented by daily rainfall data from a single weather station as a function of time, while a

multi-output (multitask) problem is represented by daily rainfall data from several stations located at different spatial positions. In the single-output case, the inter-dependencies and correlations between weather stations are neglected. This type of Gaussian process is called IGPs (Independent Gaussian Processes). IGPs are conceptually simple, computationally cheap, and easy to implement; however, they fail to account for correlations between outputs. In this work, however, the consideration of correlations between rainfall stations is key to our investigations. The extension of standard Gaussian process regression (GPR) to multi-output problems has received much attention in the literature and is still being actively investigated [9]. Herein, we first present the standard Gaussian process and then the multi-task formulation using GPs.

3.1. Standard or Single Gaussian Process

A Gaussian process is an infinite-dimensional multi-variate Gaussian. Let y(t) be the complete observation (*i.e.*, no missing function of time t), assuming y(t) follows the GP, then:

$$y(t) \sim GP(m, K) \tag{1}$$

With mean m and a covariance K parametrized by the set of hyperparameters Θ.

For a single GPs where y(t) = (y(t1), y(t2), …, y(tN)), the covariance function considers only the covariance between any two points of time t and t', then the covariance K = k(t,t'), and N the number of samples. Many functions can be used to represent the covariance between two instants of time. The covariance functions play a crucial role in GPs. They are used to enforce prior knowledge about the data in GP regression by defining what constitutes as similarity between the data points.

3.2. Coregionalization for GPs

The standard Gaussian Process (GP) formulation is typically applied to scalar functions, meaning it predicts only a single variable. When extending the standard GP formulation to multiple variables, Independent Gaussian Processes (IGPs) can be considered as an alternative. However, IGPs do not account for the correlation between outputs, which can reduce forecasting accuracy when dealing with real variables that are interdependent.

In contrast, the multi-task Gaussian process (MTGP) framework, as proposed by [9], addresses the correlation between outputs. This approach incorporates the correlation between inputs and outputs by using the linear model of coregionalization (LMC) within the GP framework. The LMC represents outputs as linear combinations of Q independent random functions, $u_q(x)$, which are modeled as Gaussian processes with zero mean and covariance k_q as their prior. Assuming M outputs represented by $\{f_m(\mathbf{x})\}_{m=1}^{M}$, f_m can be expressed as:

$$f_m = \sum_{q=1}^{Q} a_{m,q} u_q(\mathbf{x}) \tag{2}$$

where $a_{m,q}$ are scalar coefficients and $u_m \sim \mathcal{GP}\left(0, k_q(\mathbf{t},\mathbf{t})\right)$. The LMC assumes uncorrelated latent Gaussian process *i.e.* $cov\left(u_q(\mathbf{x}), u_{q'}(\mathbf{t})\right) = 0$ for $q/ \equiv q'$. Then the covariance function of two outputs function f_m and $f_{m'}$ can be expressed as:

$$cov\left(f_m(\mathbf{x}), f_{m'}(\mathbf{x})\right) = \boldsymbol{K}(\mathbf{t},\mathbf{t}') = \sum_{q=1}^{Q} B_q k_q(\mathbf{t},\mathbf{t}') \tag{3}$$

for Q = 1, we have the simplified case of LMC known as intrinsic corgeionalization matrix (ICM) then the covariance of M outputs is expressed as:

$$\boldsymbol{K} = B \otimes k(\mathbf{t},\mathbf{t}') \tag{4}$$

For the multi-task Gaussian process proposed by \cite{multitask} the covariance function considered the correlation between inputs and outputs. The dataset used for training is the input represented by time and the daily rainfall data at M weather stations. The daily rainfall at M weather stations can be expressed as:

$$\mathbf{y}(\mathbf{t}) = \left(y_1(t_1), \cdots, y_1(t_N), \cdots, y_2(t_1), \cdots, y_2(t_N), \cdots, y_M(t_1), \cdots, y_M(t_N)\right)^T$$

where y_{il} is the rainfall observation of l^{th} station at time t_i with $i = 1 \cdots N$ and $l = 1 \cdots M$. The hyperparameters are obtained using dataset and the minimization of the likelihood, the predict mean and covariance can be expressed as Equations (5) and (6).

$$m_* = K_*^T K^{-1} y. \tag{5}$$

$$C_* = K_{*,*} - K_*^T K^{-1} K_*, \tag{6}$$

The python framework for Gaussian process named GPy [15] and [16] is used for GP simulation.

4. Methodology

This section presents the methodology used to address the problem of imputing missing daily rainfall data from weather stations using Multi-Task Gaussian Processes (MTGPs) and compares it to traditional statistical methods and machine learning techniques. The steps involved in the methodology are as follows.

4.1. Data Organization and Grouping of Stations

We first organize the rainfall data by grouping the stations around synoptic stations. The synoptic stations, which have complete data over five years, serve as the central reference points for these groups. Each synoptic station (shown in red in **Figure 1** forms a group with the stations located within a 100 km radius. This grouping ensures that nearby stations, which are likely to have correlated rainfall patterns, are considered together for data imputation.

For each group, the neighbouring stations are treated as dependent variables,

and the missing data of each station are estimated based on the available data from its group members. The stations are grouped into 10 groups, with each group associated with a synoptic station. For example, the blue color in Figure 2 represents the neighbouring meteorological sensors at the Bobo synoptic station. Some synoptic stations, such as Bobo, Boromo, and Ouaga, have more neighbouring stations, while others, like Bogandé, have fewer.

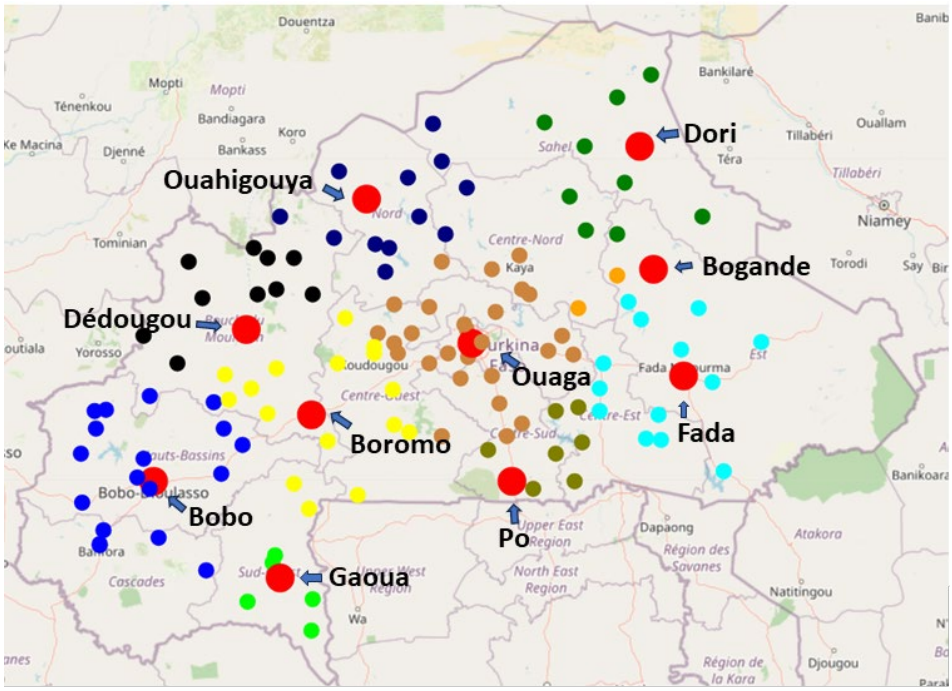

Figure 2. The 10 groups of stations built around the synoptic stations, each group represented by a color and the synoptic stations in red.

4.2. Comparison with Traditional Imputation Methods

To assess the performance of the MTGPs in imputing missing rainfall data, we compare its performance with several traditional statistical methods and machine learning techniques widely used for data imputation in the literature [1] [4]. The methods considered in this comparison include:

- **Deletion Method**: This method involves removing any records with missing values. This is a simple approach but can lead to significant data loss.
- **Mean/Mode Imputation**: Missing values are filled with the mean (for continuous data) or mode (for categorical data) of the available data for each variable.
- **Regression Techniques**: A regression model is used to predict the missing values based on the relationships between the variables.
- **K-Nearest Neighbors (KNN)**: This technique fills missing values by averaging the values of the nearest neighbors in the feature space.
- **Matrix Factorization (MF)**: Matrix factorization methods decompose the data matrix into factors and use them to predict the missing values.
- **Multiple Imputation by Chained Equations (MICE)**: MICE creates multiple

imputations for missing data and averages the results to provide a robust estimation.

4.3. Performance Evaluation

In order to evaluate the performance of different imputation methods listed in section above. Two performance indicators are used, namely the correlation coefficient ρ and the root-mean-square error (RMSE).

The comparison is conducted for each method, and the one yielding the best results in terms of accuracy and reliability is selected. In the following subsections, we present the results of these comparisons and discuss the implications for improving the imputation of missing daily rainfall data.

5. Results

In this section, we investigate the effectiveness of Multi-Task Gaussian Processes (MTGPs) for imputing daily rainfall data by utilizing correlations between nearby stations. The section is divided into three parts: first, we analyze the dataset used for the investigation; second, we explore the most suitable Gaussian process models for imputation; and third, we compare MTGPs with other imputation methods.

For our analysis, the dataset for each group is organized in a days-stations matrix form, where the columns represent data from individual stations and the rows represent data for the same day. The rainfall data for Burkina Faso from the year 2021 are used in this study. The days-stations rainfall matrices for the groups of Ouaga, Bobo, Bogandé, Fada, Gaoua, and Ouahiguya are presented in **Figures 3-5**. As shown in these figures, each group contains naturally occurring missing data, represented by white tones, and the rainfall patterns vary across groups.

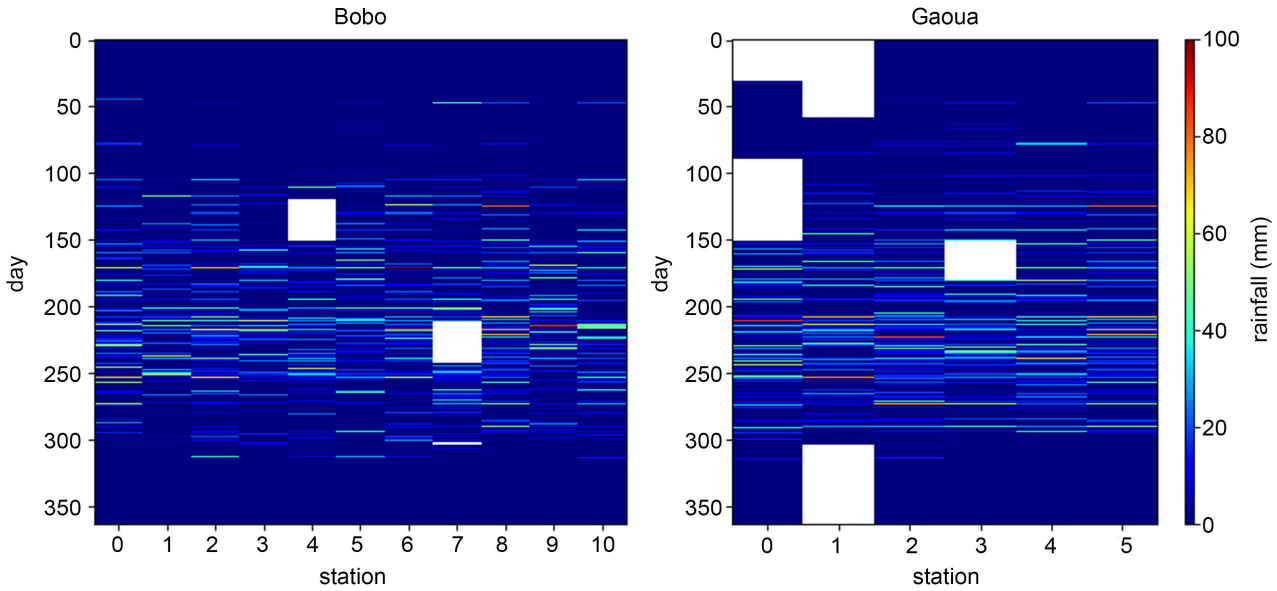

Figure 3. The rainfall matrix of days-stations group of Bobo and Gaoua.

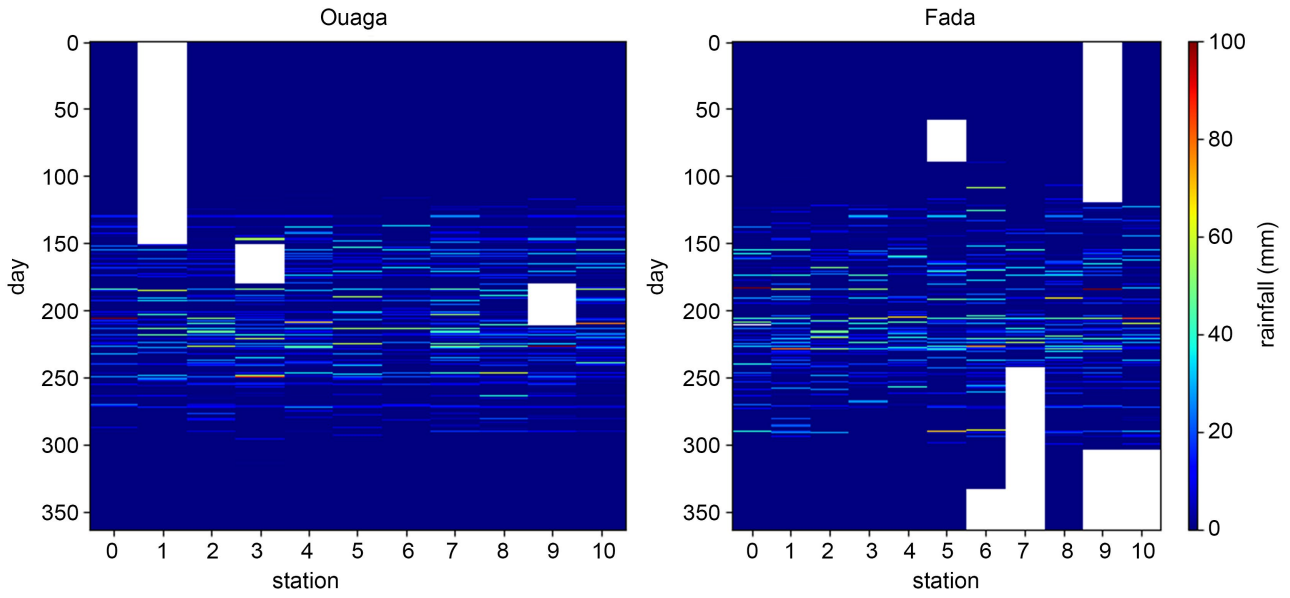

Figure 4. The rainfall matrix of days-stations group of Ouaga and Fada.

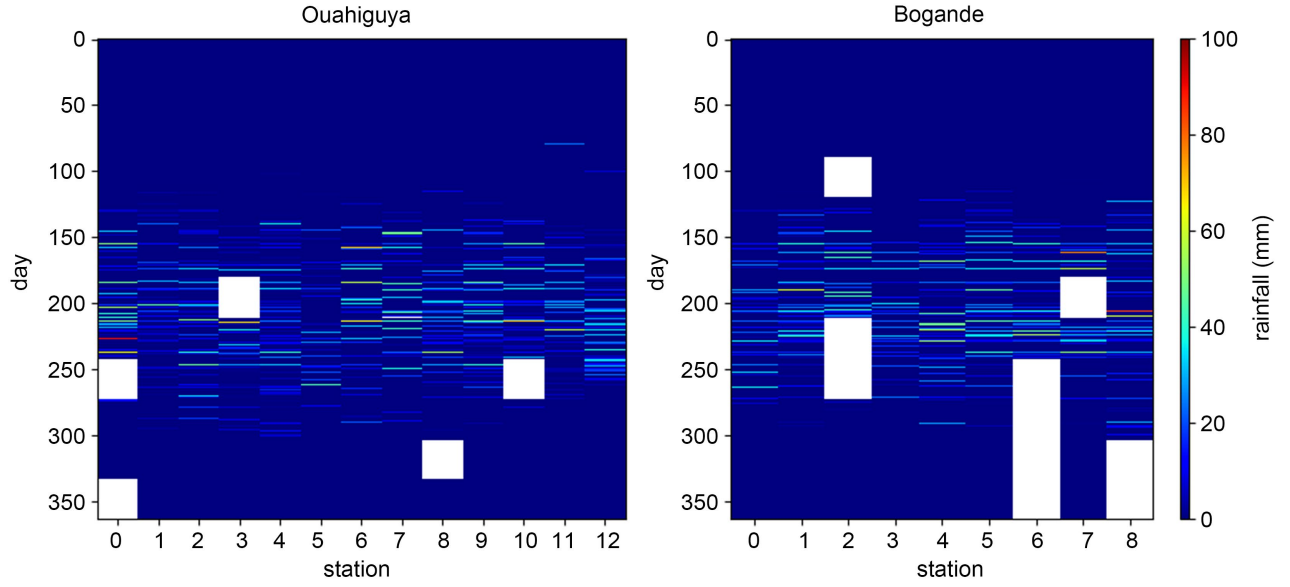

Figure 5. The rainfall matrix of days-stations group of Ouahiguya and Bogande.

In Figure 1, the Bobo and Gaoua station groups exhibit similarities. In both groups, rainfall begins early in the year (within the first 50 days), and dry or low-intensity days (shaded blue) occur roughly before the 50th day and after the 300th day of the year. This pattern can be explained by the fact that both groups are located in the Sudanese climatic zone. Similarly, in Figure 2, the Ouaga and Fada station groups show a similar pattern, with dry and rainy days occurring in roughly the same time frames, with rainy days spanning between the 100th and 300th days of the year. This pattern can be attributed to both groups being located in the Sudano-Sahelian climatic zone. In Figure 3, the rainfall intensity is lower than that

observed in Figure 1 and Figure 2, reflecting conditions in the Sahelian zone, where rain is more sporadic and concentrated within a few days. In this case, the rainy days are concentrated between the 150th and 300th days of the year. While rainfall patterns differ across climatic zones, they exhibit similarities within the same zone. For this reason, in the subsequent analysis, we focus on one group from each climatic zone: Bobo from the Sudanese zone, Ouaga from the Sudano-Sahelian zone, and Bogandé from the Sahelian zone.

5.1. Evaluation of Imputation Methods Using Artificial Missing Data for Rainfall Stations

To assess the performance of different imputation techniques, artificial missing data is introduced into the complete datasets of these stations. Missing data is simulated at levels of 20%, 40%, and 50% for stations located within a 100 km radius of the synoptic stations.

The performance of various imputation models is then evaluated and compared using the artificially generated missing data. The data is introduced on stations close to synoptic stations, defined as those within a 100 km radius.

5.2. Comparison between MTGPs and Other Imputation Methods

In the following section, we identify the best Gaussian Process (GP) model for imputing daily rainfall data. Specifically, we compare the multitask GP model (MTGP) with well-known statistical and machine learning imputation methods such as MICE, KNN, LOCF, and mean imputation (ME). The comparison is made using artificially generated missing data at levels of 20%, 40%, and 50% for daily rainfall across all stations.

Tables 1-6 present performance indicators, including RMSE and the correlation coefficient ρ, for each station in the Ouaga, Bobo, and Bogandé groups. As shown in these tables, the MTGPs model outperforms other imputation methods in each group. As the percentage of missing data increases, MTGPs continues to provide better results than traditional imputation techniques.

In the Bobo group, with 20% missing data, only two stations exhibit better performance than MTGPs. When the missing data percentage increases to 50%, only one station outperforms MTGPs. Across all stations in this group, MTGPs consistently outperforms MICE.

In the Ouaga group, MTGPs demonstrate superior performance compared to other methods. Among the 11 stations in the Ouaga group, MTGPs achieves the lowest RMSE and highest correlation coefficient (ρ) in 8 stations for 20% missing data. When the missing data rises to 50%, MTGPs achieves the best results in 10 stations. For 20% missing data, KNN is the second-best method, while for 50%, MICE emerges as the second-best performer.

In the Bogandé group, for 20% missing data, MTGPs achieve the lowest RMSE and highest ρ in 4 stations, KNN in 2 stations, and ME in 2 stations. However, for 50% missing data, MTGPs improve their performance, achieving the best results in 5 stations, while KNN outperforms in 3 stations.

Table 1. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios, estimated 20% on the group of Bobo. The indicator types are RMSE and ρ on the daily rainfall.

Name	RMSE (mm)					ρ				
	MTGPs	MICE	KNN	LOC	ME	MTGPs	MICE	KNN	LOC	ME
Banfora Agri	**3.391**	3.672	4.115	3.677	3.585	**0.942**	0.932	0.914	0.932	0.934
Bereba	**4.239**	4.569	4.46	5.772	4.575	**0.845**	0.818	0.826	0.713	0.815
Bobo-Dioulasso	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0
Bondoukuy	2.767	4.012	2.286	3.175	2.187	0.944	0.89	0.961	0.926	**0.965**
Hounde	1.32	3.927	1.22	3.674	1.972	0.985	0.895	**0.987**	0.89	0.967
Koumbia	**3.46**	5.201	3.684	4.692	3.92	**0.907**	0.792	0.894	0.827	0.879
Nasso	**2.495**	3.604	2.906	4.891	3.798	**0.973**	0.947	0.963	0.893	0.935
Orodara	**3.106**	3.884	3.578	6.37	3.493	**0.92**	0.886	0.893	0.742	0.897
Ouo	**2.876**	3.62	2.852	5.291	2.934	**0.963**	0.942	0.963	0.882	0.961
Samorogouan	**4.617**	4.832	4.855	6.822	4.822	**0.878**	0.868	0.863	0.745	0.864
Sideradougou	**3.46**	4.855	3.586	4.488	3.599	**0.907**	0.824	0.901	0.845	0.899

Table 2. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios, estimated 50% on the group of Bobo. The indicator types are RMSE and ρ on the daily rainfall.

Name	RMSE (mm)					ρ				
	MTGPs	MICE	KNN	LOC	ME	MTGP	MICE	KNN	LOC	ME
Banfora agri	**7.236**	7.896	7.941	9.79	7.901	**0.7**	0.65	0.616	0.483	0.621
Bereba	5.931	6.086	5.843	6.5	6.047	0.665	0.65	0.679	0.609	0.643
Bobo-dioulasso	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0
Bondoukuy	**5.76**	6.112	6.445	8.605	6.633	**0.72**	0.675	0.624	0.445	0.595
Hounde	**4.77**	5.532	5.356	7.664	5.493	**0.778**	0.696	0.707	0.533	0.688
Koumbia	**5.598**	8.113	6.312	6.71	6.428	**0.734**	0.481	0.641	0.616	0.623
Nasso	**3.393**	4.362	4.934	8.042	6.434	**0.949**	0.914	0.892	0.696	0.8
Orodara	**3.907**	4.921	4.849	8.119	4.687	**0.871**	0.792	0.794	0.602	0.808
Ouo	**7.246**	8.606	8.059	10.494	7.727	**0.731**	0.621	0.657	0.486	0.682
Samorogouan	**7.342**	7.796	7.738	9.338	7.699	**0.65**	0.594	0.589	0.469	0.595
Sideradougou	**5.583**	6.152	5.811	7.144	5.565	**0.739**	0.693	0.708	0.614	0.736

Table 3. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios, estimated 20% on the group of Ouaga. The indicator types are RMSE and ρ on the daily rainfall.

Name	RMSE (mm)					ρ				
	MTGPs	MICE	KNN	LOC	ME	MTGPs	MICE	KNN	LOC	ME
Bousse	**1.936**	3.817	2.391	6.984	2.689	**0.973**	0.901	0.958	0.741	0.947
Guilongou	**4.055**	4.512	4.32	4.98	4.553	**0.893**	0.869	0.879	0.837	0.863
Kindi	1.778	2.158	1.419	4.108	1.779	**0.968**	0.954	0.979	0.849	0.967
Kokologho	**2.534**	3.145	2.75	4.438	3.626	**0.958**	0.938	0.951	0.871	0.912
Kombissiri	**4.074**	5.279	4.16	4.746	4.251	**0.843**	0.749	0.836	0.79	0.827

Continued

Korsimoro	**2.936**	3.503	3.141	4.697	3.502	**0.921**	0.887	0.909	0.8	0.885
Mane	**0.954**	1.851	1.119	3.291	1.336	0.984	0.943	**0.978**	0.841	0.969
Ouagadougou aero	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0
Sapone	**3.012**	3.792	3.477	3.864	4.11	**0.903**	0.852	0.869	0.834	0.809
Tanghin dassouri	3.229	3.422	1.414	6.323	2.583	**0.936**	0.931	0.984	0.753	0.944
Toece	1.141	3.397	0.934	6.199	1.54	0.989	0.909	**0.993**	0.765	0.981

Table 4. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios, estimated 50% on the group of Ouaga. The indicator types are RMSE and ρ on the daily rainfall.

Name	RMSE (mm)					ρ				
	MTGPs	**MICE**	**KNN**	**LOC**	**ME**	**MTGPs**	**MICE**	**KNN**	**LOC**	**ME**
Bousse	**7.144**	7.43	7.549	8.48	7.647	**0.522**	0.463	0.431	0.275	0.406
Guilongou	5.818	**5.433**	6.201	9.426	6.24	0.766	**0.798**	0.733	0.47	0.721
Kindi	**3.451**	4.266	4.318	7.477	4.726	**0.867**	0.798	0.782	0.513	0.732
Kokologho	**5.84**	7.878	6.474	8.122	7.039	**0.76**	0.489	0.693	0.486	0.609
Kombissiri	**3.292**	3.821	3.407	5.184	3.962	**0.904**	0.876	0.896	0.771	0.859
Korsimoro	**5.207**	5.746	5.588	6.646	5.828	**0.728**	0.652	0.67	0.553	0.634
Mane	**2.108**	2.983	2.633	3.947	2.58	**0.922**	0.876	0.888	0.751	0.888
Ouagadougou aero	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0
Sapone	**4.112**	4.588	4.66	6.904	5.223	**0.813**	0.763	0.747	0.479	0.664
Tanghin dassouri	**2.085**	3.723	2.555	5.057	3.787	**0.965**	0.88	0.948	0.793	0.876
Toece	**5.532**	5.847	5.654	8.183	6.186	**0.697**	0.647	0.679	0.422	0.589

Table 5. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios, estimated 20 \% on the group of Bogande. The indicator types are RMSE and ρ on the daily rainfall.

Name	RMSE (mm)					ρ				
	MTGPs	**MICE**	**KNN**	**LOC**	**ME**	**MTGPs**	**MICE**	**KNN**	**LOC**	**ME**
Bani	**3.12**	4.294	3.538	4.264	3.477	**0.787**	0.578	0.711	0.59	0.722
Bogande	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0
Boulsa	**3.983**	4.07	4.449	4.999	4.473	**0.849**	0.845	0.805	0.755	0.802
Dakiri	**3.039**	3.389	3.413	3.909	3.59	**0.795**	0.75	0.733	0.638	0.7
Gayeri	3.967	6.144	3.949	6.621	**3.871**	0.836	0.648	0.836	0.62	**0.843**
Kossougoudou	**2.373**	3.738	3.412	5.052	3.754	**0.945**	0.87	0.885	0.75	0.857
Piela	1.824	3.366	**0.9**	3.341	1.502	0.974	0.92	0.993	0.915	0.982
Sebba	2.001	2.025	**1.955**	4.928	2.042	0.965	0.964	0.967	**0.823**	0.963
Yamba	3.148	5.789	2.638	2.695	**2.589**	0.93	0.803	0.95	0.948	**0.951**

Table 6. Statistical indicators describing the performance of the considered techniques with different missing-data scenarios, estimated 50% on the group of Bogande. The indicator types are RMSE and ρ on the daily rainfall

Name	RMSE (mm)					ρ				
	MTGPs	**MICE**	**KNN**	**LOC**	**ME**	**MTGPs**	**MICE**	**KNNs=**	**LOC**	**ME**
Bani	**3.276**	3.427	3.791	5.222	3.669	**0.761**	0.728	0.658	0.461	0.679
Bogande	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0
Boulsa	**5.205**	5.678	5.619	6.683	5.54	**0.727**	0.66	0.659	0.539	0.67
Dakiri	**3.39**	4.082	3.626	5.215	3.902	**0.739**	0.598	0.697	0.488	0.626
Gayeri	5.549	5.732	**5.519**	7.428	5.568	0.647	0.615	**0.649**	0.478	0.636
Kossougoudou	**3.46**	4.165	4.894	6.27	5.404	**0.882**	0.823	0.751	0.57	0.67
Piela	**4.202**	4.948	5.45	8.503	6.129	**0.847**	0.778	0.734	0.478	0.625
Sebba	3.483	4.745	**3.05**	4.323	3.612	0.902	0.835	**0.917**	0.843	0.882
Yamba	4.14	4.665	**3.865**	5.672	3.951	0.874	0.843	**0.89**	0.778	0.886

5.3. Conclusion

This study explored the efficacy of Multi-Task Gaussian Processes (MTGPs) for imputing missing daily rainfall data, leveraging correlations between nearby meteorological stations. The results demonstrated that MTGPs outperform well-established statistical and machine learning methods, such as MICE, KNN, LOCF, and mean imputation (ME), across various missing data scenarios (20%, 40%, and 50%) in different climatic zones. The superior performance of MTGPs is particularly evident as the percentage of missing data increases, underscoring their robustness and adaptability in handling data sparsity.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Chiu, P.C., Selamat, A. and Krejcar, O. (2019) Infilling Missing Rainfall and Runoff Data for Sarawak, Malaysia Using Gaussian Mixture Model Based K-Nearest Neighbor Imputation. In: Wotawa, F., Friedrich, G., Pill, I., Koitz-Hristov, R. and Ali, M., Eds., *Advances and Trends in Artificial Intelligence. From Theory to Practice*, Springer, 27-38. https://doi.org/10.1007/978-3-030-22999-3_3

[2] Teegavarapu, R.S.V. (2014) Missing Precipitation Data Estimation Using Optimal Proximity Metric-Based Imputation, Nearest-Neighbour Classification and Cluster-Based Interpolation Methods. *Hydrological Sciences Journal*, **59**, 2009-2026. https://doi.org/10.1080/02626667.2013.862334

[3] Lachin, J.M. (2015) Fallacies of Last Observation Carried Forward Analyses. *Clinical Trials*, **13**, 161-168. https://doi.org/10.1177/1740774515602688

[4] Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J. (2011) Multiple Imputation by Chained Equations: What Is It and How Does It Work? *International Journal of Methods in Psychiatric Research*, **20**, 40-49. https://doi.org/10.1002/mpr.329

[5] Oyerinde, G.T., Lawin, A.E. and Adeyeri, O.E. (2021) Multi-variate Infilling of Missing Daily Discharge Data on the Niger Basin. *Water Practice and Technology*, **16**,

961-979. https://doi.org/10.2166/wpt.2021.048

[6] Nakagawa, S. (2015) Missing data. In: Fox, G.A., Ed., *et al.*, Eds., *Ecological Statistics*, Oxford University Press, 81-105. https://doi.org/10.1093/acprof:oso/9780199672547.003.0005

[7] Sa'adi, Z., Yusop, Z., Alias, N.E., Chow, M.F., Muhammad, M.K.I., Ramli, M.W.A., *et al.* (2023) Evaluating Imputation Methods for Rainfall Data under High Variability in Johor River Basin, Malaysia. *Applied Computing and Geosciences*, **20**, Article ID: 100145. https://doi.org/10.1016/j.acags.2023.100145

[8] Vidal-Paz, J., Rodríguez-Gómez, B.A. and Orosa, J.A. (2023) A Comparison of Different Methods for Rainfall Imputation: A Galician Case Study. *Applied Sciences*, **13**, Article 12260. https://doi.org/10.3390/app132212260

[9] Hastie, T., Tibshirani, R. and Friedman, J. (2001) The Elements of Statistical Learning. Springer.

[10] Bonilla, E.V., Chai, K. and Williams, C. (2007) Multi-Task Gaussian Process Prediction. https://proceedings.neurips.cc/paper_files/paper/2007/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf

[11] Álvarez, M.A., Rosasco, L. and Lawrence, N.D. (2012) Kernels for Vector-Valued Functions: A Review. Now Foundations and Trends. https://doi.org/10.1561/9781601985590

[12] Liu, H., Cai, J. and Ong, Y. (2018) Remarks on Multi-Output Gaussian Process Regression. *Knowledge-Based Systems*, **144**, 102-121. https://doi.org/10.1016/j.knosys.2017.12.034

[13] Konomi, B., Karagiannis, G. and Lin, G. (2015) On the Bayesian Treed Multivariate Gaussian Process with Linear Model of Coregionalization. *Journal of Statistical Planning and Inference*, **157**, 1-15. https://doi.org/10.1016/j.jspi.2014.08.010

[14] Borchani, H., Varando, G., Bielza, C. and Larrañaga, P. (2015) A Survey on Multi-output Regression. *WIREs Data Mining and Knowledge Discovery*, **5**, 216-233. https://doi.org/10.1002/widm.1157

[15] (2012) GPy: A Gaussian Process Framework in Python. GPy, Sheffield Machine Learning. http://github.com/SheffieldML/GPy

[16] de Wolff, T., Cuevas, A. and Tobar, F. (2020) MOGPTK: The Multi-Output Gaussian Process Toolkit. arXiv: 2002.03471. https://arxiv.org/abs/2002.03471