Evaluation of Imputation Techniques using Reanalysis Data for Meteorological Variables in Northern Chile

Francisco García Barrera, David Contreras Aguilar, Héctor Aldea Navarro, Pablo Cárcamo Zúñiga, Mauricio Oyarzún Silva, Alonso Inostrosa-Psijas, Gabriel Icarte Ahumada, and Francisco Moreno Herrera

***Abstract*—The article explores a study on meteorological data imputation in Northern Chile, an arid region with complex geomorphology. Obtaining complete and high-quality time series poses a challenge due to data loss at meteorological stations, hindering climate change analysis in the area. Six imputation techniques were evaluated using reanalysis data from the CFSR and CFSv2 models, integrated into a single data set as an alternative to the use of neighboring meteorological stations. These models are valuable for the study of climate, especially when meteorological stations are not available or have data problems. For the research work, the six stages of the CRISP-DM methodology were developed, providing a robust framework. The results show that the Direct Imputation, Hot-Deck, Weighted K-Nearest Neighbor Imputation and Inverse Distance Weighting techniques obtain the lowest residual errors according to meteorological variables, while the NR technique is consistently inferior compared to the other techniques evaluated. The study concludes that it is essential to evaluate imputation techniques and reanalysis models based on the specific geographic area where they will be applied. Reanalysis data represents the study area's behavior and meteorological variables with varying degrees of accuracy. As a result, the best imputation technique differs depending on the geographic region, reanalysis model, and meteorological variable.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10340

***Index Terms*—Data Mining, Imputation, Meteorological Data, Northern of Chile, Reanalysis Data**

I. Introduction and Background

The associate editor coordinating the review of this manuscript and approving it for publication was Anabel Martin (*Corresponding author: Francisco García Barrera*).

This work was supported by the National Agency for Research and Development (ANID), Chile, under the Fondecyt de Iniciación grant number 11230961, awarded to Alonso Inostrosa-Psijas.

Francisco García Barrera, H. A. Navarro, M. O. Silva, and G. I. Ahumada are with the Faculty of Engineering and Architecture, Universidad Arturo Prat, Iquique 1110939, Tarapacá, Chile (e-mails: francgar@unap.cl, haldea@estudiantesunap.cl, moyarzunsil@unap.cl, and gicarte@unap.cl).

D. C. Aguilar is with the Smart Society Research Group, La Salle-Universitat Ramon Llull, Barcelona 08022, Spain (e-mail: david.contreras@salle.url.edu).

P. C. Zuñiga is with the Center for Teaching Development and Innovation, Universidad Católica de la Santísima Concepción, Concepción 4090541, Biobío, Chile (e-mail: pcarcamo@ucsc.cl).

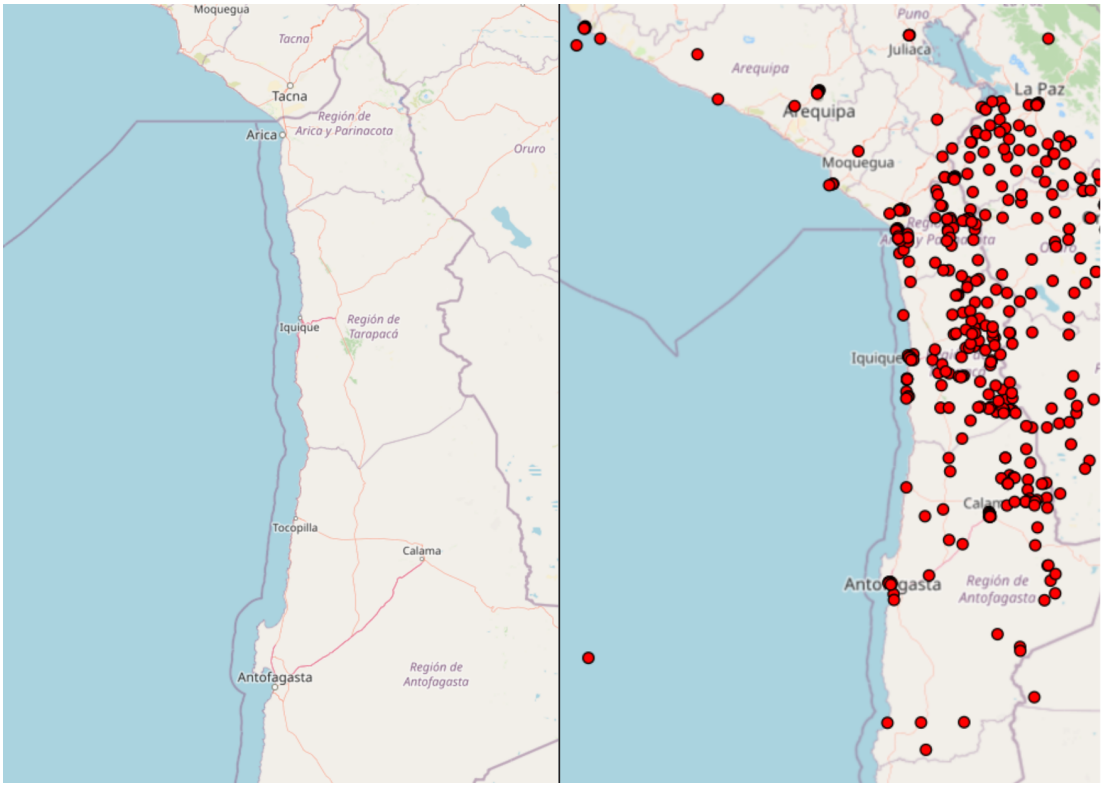
A. I. Psijas is with the School of Computer Engineering, Universidad de Valparaíso, Valparaíso 2362905, Chile (e-mail: alonso.inostrosa@uv.cl).

F. M. Herrera is with the Department of Mathematics and Computer Science, Universidad de Santiago de Chile, Santiago 9170022, RM, Chile (e-mail: francisco.moreno@usach.cl).

In recent decades, climate research has advanced substantially thanks to reanalysis models that combine historical observations with advanced numerical simulations, providing an integrated and consistent representation of climate variability over time [1], [2]. Among widely used models are the Climate Forecast System Reanalysis (CFSR) and its successor, the Climate Forecast System Version 2 (CFSv2), developed by the National Centers for Environmental Prediction (NCEP) [2], [3]. Since 1979, these models have offered continuous and homogeneous climate fields that are particularly valuable in regions with scarce or discontinuous in-situ observations [2]. This is the case of northern Chile (Fig. 1a), a climatically extreme region of interest for climate-change assessment, where meteorological stations (MS) frequently exhibit interruptions that hinder the construction of continuous time series (TS) and the estimation of Climate Extreme Indicators (CEI) [3]. In this context, reanalysis models can support gap-filling and subsequent analyses of variability and extremes [4].

Missing-data imputation is therefore a key step in climate data analysis, as it enables the reconstruction of incomplete TS and improves the reliability of downstream indicators. Prior work has addressed temporal gaps using a variety of techniques, often leveraging information from neighboring stations and/or the historical behavior of the target series [5], [6], [7], [8]. For instance, [5] evaluated multiple approaches for daily precipitation in the Soummam watershet, Algeria (e.g., k-nearest neighbors, Hot Deck, regression-based methods, and multiple imputation) under different missingness scenarios, reporting that station-based similarity can be effective and that hybrid strategies may further improve accuracy. Likewise, [6] applied the normal ratio and inverse distance weighting to complete daily precipitation records in the Valle del Cauca, Colombia, showing that distance-based methods benefit from using several surrounding stations when available.

Beyond station-to-station strategies, reanalysis models have also been considered as alternative sources to complete meteorological TS, especially where station coverage is limited or gaps are widespread. An analysis of uncertainties between reanalysis data and ground measurements, using reanalysis models such as ERA-Interim and NASA's satellite data [9], concluded that reanalysis can provide a robust basis for completing missing observations, although discrepancies with in-situ measurements introduce uncertainties that should be properly characterized. In addition, multiple imputation [10]



(a) Great Northern Chile (b) Initial MS collected

Fig. 1. Study area and distribution of meteorological stations used in the analysis.

provides a general framework that generates several plausible completed datasets and combines the results, improving the robustness when there are long or systematic gaps.

Despite these advances, two practical limitations remain particularly relevant for arid and sparsely monitored regions such as northern Chile. First, approaches relying on neighboring MS can be constrained by low station density and the fact that nearby series may also contain substantial missingness. Second, the behavior of imputation methods under extreme climatic conditions is less documented, although geographic and meteorological factors can influence imputation performance [5], [6]. To address these issues, this study evaluates the imputation of meteorological data using CFSR/CFSv2 reanalysis fields as an alternative source of information to neighboring meteorological stations, thereby reducing the dependency on nearby station records.

The goal of this research is to identify techniques that yield residual errors closest to zero, indicating the most suitable methods for reconstructing missing observations. To this end, we generate both continuous and random temporal gaps over multiple meteorological variables and evaluate several imputation techniques under controlled missingness scenarios.

The main contributions of this work are fourfold: (i) we validate and compare multiple imputation techniques using reanalysis models (CFSR and CFSv2) instead of neighboring MS data, mitigating the risk that reference stations are also incomplete; (ii) we demonstrate how reanalysis models can improve the quality and continuity of meteorological TS in regions with limited observational coverage, facilitating the computation of CEI and other indicators for climate-change assessment; (iii) we provide a methodological framework transferable to other geographic contexts and disciplines requiring complete and coherent TS; and (iv) we report empirical insights into the behavior of imputation methods in one of the most arid regions worldwide, offering evidence useful for climate research, mitigation, and adaptation efforts.

II. Materials and Methods

A. Study Area, Time Series, and Reanalysis Data

The study area covers great northern Chile with 185,148.2 km2 (Arica and Parinacota, Tarapacá, and Antofagasta regions), featuring five major geomorphological units: Coastal Plain, Coastal Range, Intermediate Depression, Altiplano, and Andes Mountain Range. These units, running longitudinally from the Pacific to the borders with Bolivia and Argentina, create a landscape of extreme geographical contrasts. The climate of the region is predominantly arid, with specific subtypes: Cloudy Coastal Desert, Inland Desert, High Altitude Marginal Desert, and HighAltitude Steppe climates. Fig. 1a presents the study area corresponding to great northern Chile, while Fig. 1b illustrates the initially collected MS, which were filtered according to data quality and the study area.

Due to the extreme climate of the region and its importance under climate change, this research focuses on TS of key meteorological variables: precipitation (PP), minimum (MINT), average (AVGT), and maximum (MAXT) temperatures.

A time series (TS) is a collection of recorded observations at successive intervals, essential for identifying trends and vari-

ation in climate variables like temperature and precipitation [11]. However, MS data often has gaps due to technical failures or interruptions. To address this, reanalysis data are used.

Reanalysis data reconstruct atmospheric and oceanic conditions retrospectively, integrating information from different sources (MS and satellites) through data assimilation techniques. This creates coherent datasets, filling observation gaps and enhancing climate system understanding [1].

B. Methodologies and Imputations Methods

The CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) is currently widely used in data mining compared to other methodologies or frameworks due to its robust and systematic structure [12]. It consists of six key phases: understanding the problem, understanding the data, data preparation, modeling, evaluation, and deployment [12]. In this study, the modeling phase is critical, as it is where the data imputation experiment is conducted, a crucial step in ensuring the integrity of subsequent analyses.

Within the CRISP-DM framework, data understanding and data preparation phases were devoted to assessing data quality, characterizing missingness patterns, and harmonizing meteorological station data with reanalysis sources (CFSR and CFSv2). These phases ensured that all datasets were properly structured and suitable for the application of imputation techniques.

The modeling phase was then used to implement and execute the imputation techniques as controlled experiments, where different methods were applied and parameterized over the prepared datasets. In this context, modeling does not refer to predictive modeling, but rather to the experimental execution of imputation methods whose outcomes are subsequently assessed in the evaluation phase.

In this study, reanalysis TS from CFSR and CFSv2 are used as an alternative source of information to neighboring MS for the imputation of missing values in meteorological TS. This approach is particularly relevant in regions where neighboring MS do not exist, are sparsely distributed, or present data gaps during the same periods requiring imputation. Under this framework, the target of the imputation process is always the MS TS, while the reanalysis TS are used as auxiliary input to the imputation methods. Specifically, reanalysis data provide the external information required by the imputation techniques to compute missing observations in the MS TS, effectively replacing the role traditionally assigned to neighboring stations when such observational support is unavailable or incomplete.

Data imputation is the process of assigning values to observations with data gaps, based on the available information in the sample. This issue is common in meteorological TS, where MS records are often incomplete for various reasons. In this context, several imputation methods are essential for maintaining data integrity and ensuring valid analyses. Next, we describe the imputation methods used.

Direct Imputation (DI). It is a simple strategy that involves replacing data gaps with mathematically valid values [13], [14]. In the context of this study, let x_t denote a meteorological TS from a MS, where $t = 1, \ldots, T$, and let $\mathcal{M} \subset \{1, \ldots, T\}$ be the set of time indices with missing observations. Under the DI approach, each missing value x_t with $t \in \mathcal{M}$ is replaced by the corresponding value from the aligned reanalysis TS r_t, such that:

$$\hat{x}_t = r_t, \quad \forall t \in \mathcal{M} \tag{1}$$

where $\hat{x}_t$ is the imputed value and r_t represents the reanalysis value (CFSR/CFSv2) temporally aligned with the MS TS. This formulation explicitly uses reanalysis data as auxiliary input to replace the missing values.

Hot-Deck (HD). This imputation technique assigns values to incomplete records by using similar observations within the same dataset. The process involves identifying incomplete records, searching for similar complete records, selecting one or more as donors, and using their values to fill temporary gaps [5]. In this study, for each missing value x_t with $t \in \mathcal{M}$, a donor index $s \notin \mathcal{M}$ is selected such that the corresponding reanalysis values satisfy a similarity criterion:

$$s = \arg\min_{k \notin \mathcal{M}} |r_t - r_k| \tag{2}$$

The missing value is then imputed as:

$$\hat{x}_t = x_s \tag{3}$$

where r_t denotes the reanalysis value (CFSR/CFSv2) at time t, which is always available, even when the corresponding MS observation is missing. The variable r_k represents reanalysis values at time indices $k \notin \mathcal{M}$, that is, time instants for which MS observations are available. This deterministic variant selects the donor whose reanalysis value r_k is closest to r_t, thereby ensuring that the imputed MS value is consistent with similar large-scale atmospheric conditions.

K-Nearest Neighbors Imputation (KNNI). This is an efficient method for imputing time gaps based on proximity between instances in a dataset. The method identifies the k nearest neighbors of the instance with data gaps and uses the mode for categorical variables and the mean for numerical variables. Unlike model-based algorithms, KNNI does not build a global estimator but searches for the nearest neighbors to impute each incomplete instance [15]. The KNNI is an extension of the Nearest Neighbor (NN) method, a variant of HD, which introduces the parameter k to prevent overfitting.

The KNNI technique is formally defined as follows: Given the values $(X_i, Y_i, 0)$ and the set of k nearest neighbours represented as $D_k = \{(X_j, Y_j, 1) \mid j = 1, 2, \ldots, k\}$, the imputed value $\hat{Y}$ is computed using a type-dependent aggregation rule:

$$\hat{Y} = \begin{cases} \arg\max\limits_{v \in \mathcal{V}} \left\{ \sum\limits_{(X_j, Y_j, 1) \in D_k} \mathbf{1}(Y_j = v) \right\}, & (Y \in \mathcal{C}), \\ \dfrac{1}{k} \sum\limits_{j=1}^{k} Y_j, & (Y \in \mathcal{R}), \end{cases} \tag{4}$$

where $\mathcal{C}$ denotes the set of possible categorical values for the target variable, and $\mathbb{R}$ denotes the real-valued domain (numerical target), $\mathcal{V}$ is the set of possible categorical values for the target feature, and $\mathbf{1}(Y_j = v)$ is an indicator function that returns 1 if the argument is true and 0 otherwise. Although the KNNI estimator is simple, two main challenges exist:

determining the optimal value of k and selecting the nearest neighbors.

Weighted-Nearest-Neighbors Imputation (WKNNI). This method is an estimation technique that calculates weighting coefficients based on similar TS. In the context of climate data, Euclidean distances between similar stations and the reference stations are used to compute these weight aiming to provide an accurate estimation of the data gaps. The final estimate is determined through a weighted average of neigh boring data, where closer neighbors areas signed higher weights and more distant ones are given lower weights. This is expressed by Equation (5).

$$P_x = \frac{\sum_{i=1}^{k}(P_i \cdot W_i)}{\sum_{i=1}^{k} W_i}\ ; \quad W_i = d_{xi}^{-\kappa}. \tag{5}$$

where P_x represents the estimated value for the reference station x, k is the number of neighboring stations, P_i are the observed values at neighboring stations, and W_i is the weighting coefficient, calculated as $d_{xi}^{-\kappa}$, where d_{xi} is the Euclidean distance between neighboring station i and reference station x. The friction distance κ typically ranges from 1.0 to 6.0 [5].

Inverse Distance Weighting (IDW). The method establishes that the influence of a variable at a station, for its calculation at any point, is inversely proportional to the distance between the two points [6]. The formula for calculating this influence is expressed as:

$$P_x = \frac{\sum_{i=1}^{N} \frac{1}{d_i^2} P_i}{\sum_{i=1}^{N} \frac{1}{d_i^2}}. \tag{6}$$

where P_x is the estimated value of the incomplete station, P_i are the values of the reference variable used for estimation, d_i is the distance from each reference point to the point being estimated, and N is the number of reference stations. According to [6], it is suggested to use the four nearest auxiliary stations, with each located in one of the quadrants defined by the coordinate axes that pass through the incomplete station, typically north–south and east–west.

Normal Ratio (NR). This technique involves estimating the data gaps, x(t), in a TS using information from three nearby and contemporaneous MS that are highly correlated with the series in question [16]. The formula accompanying this technique can be seen in Equation (7).

$$x(t) = \frac{1}{3}\left[\frac{\bar{x}}{\bar{x}_1}x_1(t) + \frac{\bar{x}}{\bar{x}_2}x_2(t) + \frac{\bar{x}}{\bar{x}_3}x_3(t)\right]. \tag{7}$$

where $\bar{x}$, $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$ are the means of the variables in question for the incomplete TS and the three-neighboring series, respectively; and x, $x_1(t)$, $x_2(t)$, $x_3(t)$ are the corresponding data for the neighboring series. This method takes advantage of the variability recorded at the other stations and the proportional relationship between them. Using three stations reduces the impact of potential errors at a single MS.

A comprehensive overview of the proposed methodology is presented in Fig. 2, which illustrates the complete workflow followed in this study. **Phase 1.** The process begins with the individual preprocessing of each MS, where all TS are homogenized to ensure a consistent data structure across stations. The meteorological observations are analyzed at a daily sampling frequency; however, in cases where MS originally reported data at an hourly resolution, these records are aggregated to daily values to guarantee temporal uniformity across all TS. Additionally, MS located outside the study area are discarded. **Phase 2.** A common temporal window is defined in order to maximize the number of MS with available data, excluding those that do not record the variables of interest or lack sufficient observations within the selected period. **Phase 3**. A data quality assessment is then performed on the MS TS, including noise filtering and outlier detection using Tukey's test based on interquartile ranges, complemented by physical plausibility rules specific to each variable (e.g., non-negative precipitation). Detected anomalies are carefully reviewed to distinguish measurement errors from genuine extreme events, retaining physically consistent values. **Phase 4.** Subsequently, MS are grouped independently for each meteorological variable (AVGT, MINT, MAXT, and PP) and by geographic region, prioritizing geomorphological coherence and positive inter-station correlations, while discarding stations with weak or negative correlations or lacking spatial consistency. **Phase 5.** In parallel, reanalysis TS are extracted from the CFSR and CFSv2 models using the geographic coordinates of neighboring MS and merged into a single dataset, as CFSv2 represents the temporal continuation of CFSR. **Phase 6.** These reanalysis data undergo the same quality control and noise-filtering procedures applied to the observational data. **Phase 7.** Prior to imputation, a correlation analysis between MS TS and reanalysis TS is conducted to verify positive correlations and ensure climatic consistency. **Phase 8.** Artificial temporal gaps are generated in the MS TS by considering both continuous and random missing patterns, with a progressive increase in the number of missing values from one to fifteen. **Phase 9.** Corresponding to the imputation stage, is performed iteratively together with the gap generation process: for each iteration, missing values are introduced and subsequently estimated using six imputation techniques (DI, HD, KNNI, WKNNI, IDW, and NR), treating each iteration as an independent experiment and using the reanalysis TS as auxiliary information to impute MS TS. **Phase 10.** Finally, the imputed results are evaluated through a post-imputation correlation analysis to confirm the preservation of temporal and statistical behavior, together with a residual-error analysis comparing the imputed values against the original removed observations. Imputation techniques yielding lower residual errors are considered to provide more accurate reconstructions and, therefore, better imputation performance under each experimental scenario.

III. Results

A. Experimental Setup-Data

The database used consists of 2,665,137 records from 638 MS in Northern Chile. After pre-filtering, 586 MS were selected for further processing. These MS belong to 14 organizations, both governmental and private, and cover data from 1935 to 2014, though not consistently across all years.

Although the majority of the records correspond to daily observations, some MS reported measurements at an hourly resolution, which were subsequently aggregated to daily values to ensure temporal homogeneity across all TS. For example, in the region of Antofagasta one of the MS owned by NOAA contains the oldest data, while an MS in Tarapacá has more recent records. On average, each MS has 16.3 years of data.

This study analyzed PP, MINT, AVGT, and MAXT variables. Tests were conducted to determine optimal date ranges with sufficient data for experiments. A Microsoft SQL Server database was used to consolidate MS data, processed using KNIME Analytics Platform. An automated process in KNIME filtered and selected relevant MS, while the imputation methods were implemented in Python. The complete preprocessing workflow, including the noise-filtering procedures applied prior to the experiments, is described in Section II, Materials and Methods, Subsection B (Methodologies and Imputation Methods), and summarized in Fig. 2. Specifically, noise-filtering is an integral component of the data quality analysis performed in Phases 3 and 6 of the workflow, which correspond to the quality assessment and filtering of MS data and reanalysis data, respectively.

B. Experimental Activity

To carry out the experimental with all the meteorological variables of interest (MINT, MAXT, AVGT AND PP), 4 groups of MS were selected for each variable, based on their correlation and geographic location. Daily data was used, and 15 iterations were performed to generate temporal gaps in the original TS of the MS, as it is recommended not to exceed 5% [17]. Below, an example of the process used for the continuous gap experiments in the meteorological variable MAXT and MINT respectively is described.

First, when the first iteration is executed, a continuous temporary gap is generated by eliminating 1 record, in the second iteration 2 records are eliminated, and so on. As the iterations progress, an amount of empty data equivalent to the iteration is generated. Fig. 3a shows the TS with 13 data gaps corresponding to the thirteenth iteration using the KNNI technique.

Subsequently, the data are then imputed using the selected techniques, which use the reanalysis data from the CFSR and CFSv2 models (as a single integrated dataset) corresponding to the coordinates of the neighboring MS instead of them, to evaluate the data of these models as an alternative source of information in the imputation processes, since neighboring MS are not immune to the problems of time gaps. As an example, Fig. 3a shows the original TS with artificially generated missing values for a given iteration, while Fig. 3b presents the corresponding imputed results obtained using the KNNI technique. In this case, the numerical formulation of KNNI described in Equation (4) is applied, where the missing values are estimated as the average of the k nearest neighbors.

C. Analysis of Results

The results are organized into two main sections: residual error analysis and correlation analysis. Each section examines

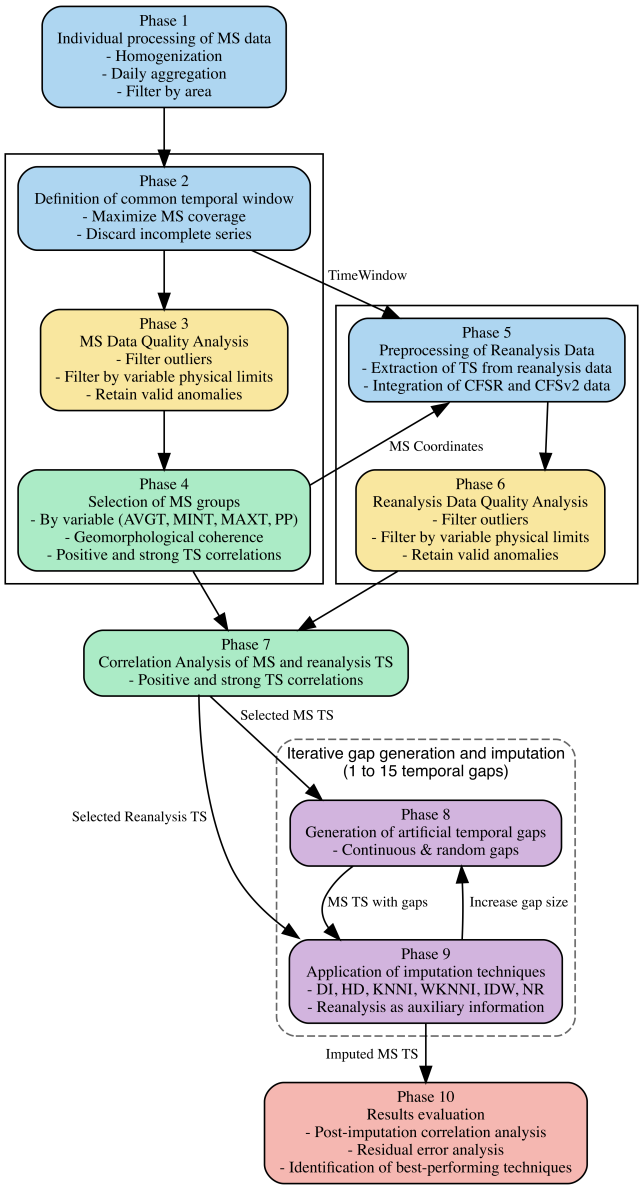
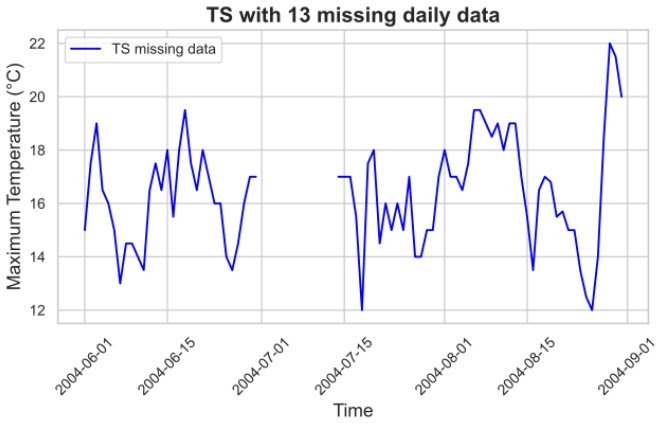


Fig. 2. Workflow of the methodology implemented in KNIME Analytics Platform and Python.

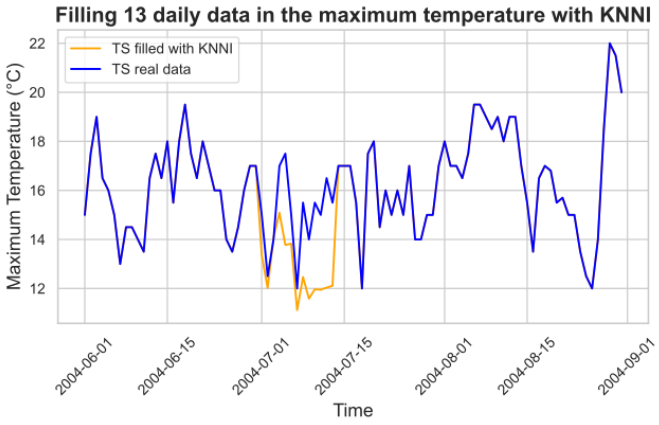
the performance of the imputation techniques across the meteorological variables (MINT, AVGT, MAXT, and PP) for both continuous and random gaps. In the residual error analysis, we evaluate the accuracy of the imputation methods by comparing the imputed values with the real data. In the correlation analysis, we use the Spearman and Kendall correlation coefficients to assess the consistency between the imputed and original time series. This structure allows for a clear understanding of how different techniques perform under varying conditions.

1) Analysis by Residual Error.:

a) Result Analysis by Continuous Temporary Gaps: The analysis of continuous gaps for the MINT variable (see Fig. 4a), shows that the most effective imputation methods in terms of residual error are the DI and WKNNI techniques. These methods yield the lowest residual errors, indicating high accuracy in imputing data gaps. On the other hand,



(a)



(b)

Fig. 3. Example of time gaps (a) and data generated by the KNNI technique compared to the original MAXT values for continuous time gaps (b).

the NR technique consistently exhibits the highest residual errors, making it the least effective for MINT imputation. For the AVGT variable (see Fig. 4b), the KNNI and WKNNI methods demonstrate the lowest residual errors, with WKNNI performing slightly better. These methods provide the most reliable results in filling continuous gaps in AVGT, while NR continues to show the highest residual errors, further confirming its unsuitability for this variable. In the case of MAXT, Fig. 4c shows that the HD and IDW methods perform best, with the lowest residual errors across the continuous gaps experiments. This suggests that HD is highly effective for MAXT imputation. Conversely, the DI method shows higher residual errors for continuous gaps, making it a less favorable option for this variable. Finally, for the PP variable (see Fig. 4d), the IDW technique consistently produces the lowest residual errors, confirming its superiority for imputing continuous gaps in precipitation data. The NR method, once again, shows the highest residual errors, performing poorly compared to the other techniques.

b) Result Analysis by Random Temporary Gaps: For random gaps in MINT (see Fig. 5a), the HD and IDW methods yield the best results, with both showing low residual errors and high consistency across experiments. The NR technique continues to underperform, with the highest residual errors, indicating its inadequacy for this type of gap. In the case of random gaps for AVGT, Fig. 5b reveals that WKNNI again achieves superior performance, with the lowest residual errors, followed closely by KNNI. These methods are the most reliable for imputing random gaps in AVGT. The NR technique remains the least effective, with the highest residual errors in this category as well. For MAXT (see Fig. 5c), HD continues to outperform the other methods in random gaps scenarios. The NR and Direct Imputation methods perform poorly, showing higher residual errors and making them less suitable for random gaps imputation. In the case of PP (see Fig. 5d), the IDW method performs best for random gaps, maintaining low residual errors similar to its performance with continuous gaps. The NR method presents the highest residual errors, making it the least effective method for PP imputation.

TABLE I
BEST PERFORMING IMPUTATION TECHNIQUES BY VARIABLE AND GAP TYPE

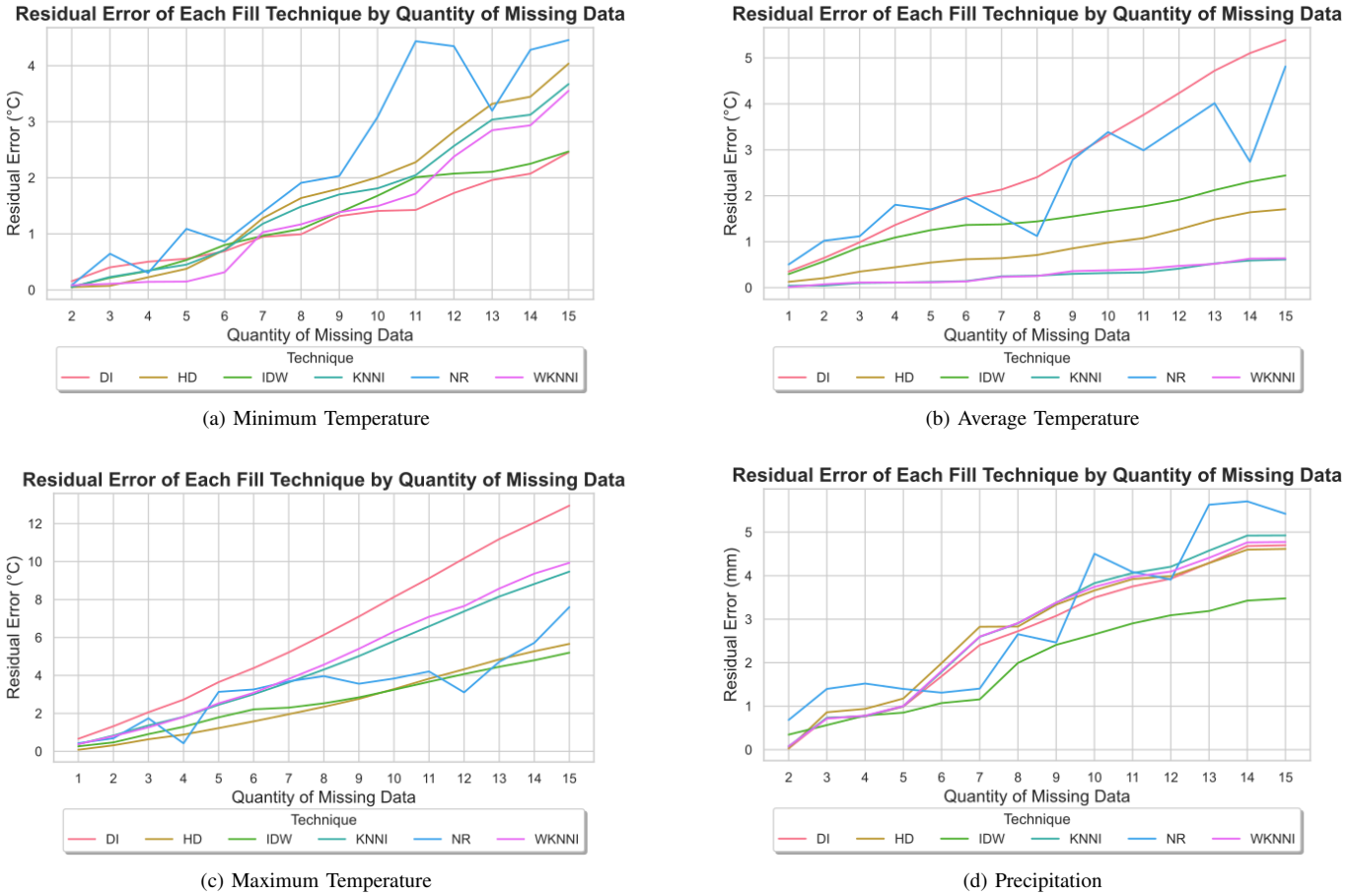
Variable	Continuous Gaps (Best)	Random Gaps (Best)
MINT (Min Temp)	DI, WKNNI	HD, IDW
AVGT (Avg Temp)	WKNNI	WKNNI, KNNI
MAXT (Max Temp)	HD, IDW	HD
PP (Precipitation)	IDW	IDW

2) Correlation Analysis (Spearman and Kendall):

a) Result analysis by continuous temporary gaps: The results show that, in the case of MINT, the DI and WKNNI methods have the highest positive correlations, indicating that the imputed data maintains a strong relationship with the real time series. When considering AVGT and MAXT, WKNNI and HD demonstrate the highest correlations. In the case of PP, the best result is obtained by IDW, while the NR method consistently presents the lowest correlations across all variables, showing weak alignment with the true data.

b) Result analysis by random temporary gaps: With random gaps, the correlation analysis reveals similar trends. In MINT, HD and WKNNI achieve the highest Spearman and Kendall correlations, while NR continues to underperform with the lowest correlations. In the case of AVGT, WKNNI stands out again, followed by KNNI, both showing strong positive correlations. In MAXT, HD demonstrates the highest correlations, confirming its effectiveness in both residual error and correlation measures. For PP, IDW maintains the strongest correlations, while NR remains the weakest performer.

To consolidate the results before presenting the conclusions, it is useful to summarize the main findings across variables and gap types (see Table I). The analysis shows that no single imputation technique is universally optimal; instead, performance depends strongly on the meteorological variable and the type of missing data scenario. Techniques such as WKNNI and HD stand out for temperature-related variables, while IDW consistently delivers the best results for precipitation. In contrast, the Normal Ratio (NR) method performs poorly across all evaluated scenarios. The results reported in Table I were obtained by simulating a total of 180 experimental



(a) Minimum Temperature

(b) Average Temperature

(c) Maximum Temperature

(d) Precipitation

Fig. 4. Residual error means according to imputation methods by continuous temporary gaps.

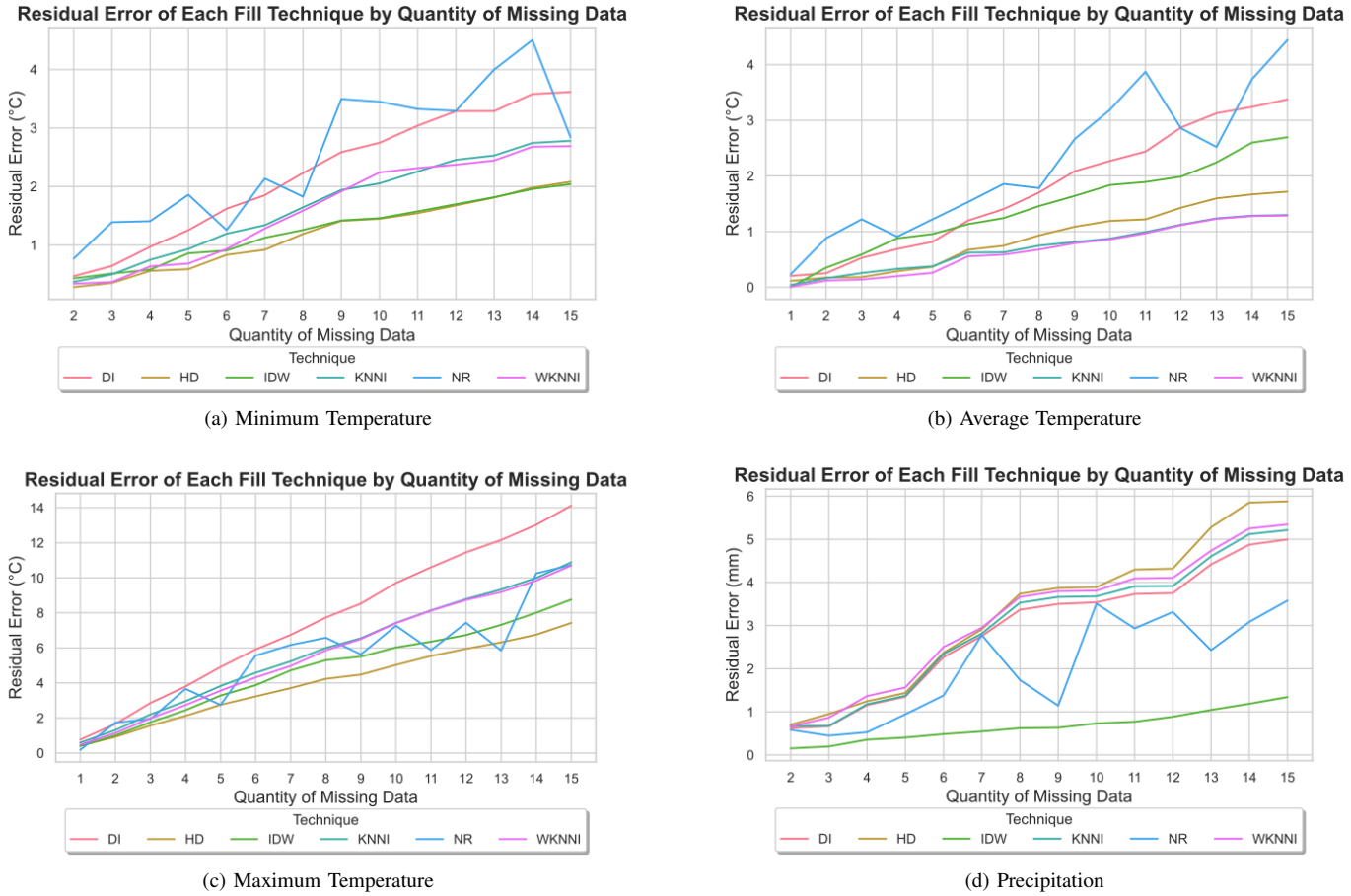
scenarios per meteorological variable. Specifically, six imputation techniques were evaluated under 15 iterative experiments, where each iteration corresponds to an experiment in which an increasing number of missing values was introduced into the original TS (from 1 missing value in the first iteration up to 15 in the last iteration). For each iteration, all six imputation techniques were applied independently. This experimental procedure was conducted under two distinct missing-data scenarios: (i) continuous temporal gaps and (ii) randomly distributed temporal gaps.

IV. Conclusions

This study demonstrates that reanalysis data can serve as a viable source for imputing meteorological variables in Northern Chile, showing strong agreement with measured records and competitive performance that depends on both the gap type and the target variable. Overall, the most reliable pairings observed were: (i) for **MINT**, DI is preferred for *continuous* gaps, whereas HD is preferred for *random* gaps; (ii) for **AVGT**, WKNNI consistently delivers the best performance under both gap types; (iii) for **MAXT**, HD achieves the best performance across all experiments, while DI and NR tend to underperform (with DI being particularly weak for continuous gaps); and (iv) for **PP**, IDW is the most effective technique for both continuous and random gaps. Across all variables and scenarios, NR is consistently inferior to the other evaluated techniques.

To make the results directly actionable, we recommend prioritizing techniques according to the missingness scenario. For *continuous gaps* (long missing segments), methods that better preserve medium-to-long temporal structure should be preferred; based on residual errors closest to zero and the consistency observed across experiments, the recommended prioritization is: **DI** for MINT, **WKNNI** for AVGT, **HD** for MAXT, and **IDW** for PP. In contrast, **NR** should be avoided under continuous gaps due to systematically larger deviations. For *random gaps* (isolated missing points), local consistency becomes dominant; accordingly, we recommend **HD** for MINT, **WKNNI** for AVGT, **HD** for MAXT, and **IDW** for PP as the first choices. When a conservative selection is required (e.g., heterogeneous datasets or mixed gap patterns), **WKNNI** (temperature-related variables) and **IDW** (precipitation) provide the most robust choices, as they remain within the top-performing group across gap types.

Beyond the numerical results, the findings reveal that the performance of each method is conditioned by the nature of the meteorological variable and the type of gap considered. This highlights the importance of tailoring the choice of imputation method to the specific analytical scenario, rather than applying a one-size-fits-all approach. Importantly, the



(a) Minimum Temperature

(b) Average Temperature

(c) Maximum Temperature

(d) Precipitation

Fig. 5. Residual error means according to imputation methods by random temporary gaps.

consistency of IDW for precipitation and HD for maximum temperature suggests that certain variable–technique pairings may be transferable to other arid or data-sparse regions, while the robust behavior of WKNNI underscores the potential of similarity-based methods when applied to temperature-related variables.

From a methodological standpoint, this work validates the use of the CRISP-DM framework for structuring the entire process, from data understanding to evaluation, in meteorological contexts. The integration of reanalysis datasets with classical imputation methods provides an alternative to relying exclusively on neighboring meteorological stations, which in Northern Chile are often scarce or affected by data interruptions. This approach contributes to more reliable long-term climate series, which are essential for assessing CEI and supporting adaptation strategies in one of the most arid regions in the world.

There are, however, limitations that should be acknowledged. The analysis relied on CFSR and CFSv2 reanalysis data, whose spatial resolution may not fully capture local-scale variability, especially in complex topographies. In addition, only classical imputation methods were evaluated; emerging approaches such as ensemble models, neural networks, or case-based reasoning could potentially reduce residual errors further and should be explored in future work. Expanding the study to other climatic regions would also help test the generalizability of the observed method–variable associations.

In summary, this study not only identifies scenario-specific, high-performing imputation techniques for distinct meteorological variables in Northern Chile, but also demonstrates the broader utility of combining reanalysis data with systematic imputation strategies. The proposed prioritization by gap type provides a practical reference for researchers and practitioners working with incomplete time series in arid and data-scarce regions, supporting more reliable climate indicators and analyses. Finally, these findings lay the groundwork for integrating advanced data-driven approaches (e.g., ensembles or learning-based models) on top of reanalysis-informed imputation to further enhance climate analysis and decision-making under global change.

ACKNOWLEDGMENTS

This work was supported by the National Agency for Research and Development (ANID), Chile, under the Fondecyt de Iniciación grant number 11230961, awarded to Alonso Inostrosa-Psijas.

REFERENCES

[1] S. M. Uppala, P. W. Kallberg, A. J. Simmons, U. Andrae, V. Da Costa Bechtold, M. Fiorino, J. K. Gibson, A. Haseler, A. Hernandez, G. A.

Kelly, X. Li, K. Onogi, E. Saarinen, N. Sokka, R. P. Allan, E. Andersson, K. Arpe, M. A. Balmaseda, A. C. M. Beljaars, L. Berg, J. Bidlot, J. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Hólm, B. J. Hoskins, L. Isaksen, P. A. E. M. Janssen, R. Jenne, A. P. McNally, J.-F. Mahfouf, J.-J. Morcrette, N. A. Rayner, R. W. Saunders, P. Simon, A. Sterl, K. E. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo, and J. Woollen, "The ERA-40 reanalysis," Q. J. R. Meteorol. Soc., vol. 131, no. 612, pp. 2961–3012, Oct. 2005, doi: 10.1256/qj.04.176.

[2] S. Saha, S. Moorthi, H.-L. Pan, X. Wu, J. Wang, S. Nadiga, P. Tripp, R. Kistler, J. Woollen, D. Behringer, H. Liu, D. Stokes, R. Grumbine, G. Gayno, J. Wang, Y.-T. Hou, H.-Y. Chuang, H.-M. H. Juang, J. Sela, M. Iredell, R. Treadon, D. Kleist, P. Van Delst, D. Keyser, J. Derber, M. Ek, J. Meng, H. Wei, R. Yang, S. Lord, H. Van Den Dool, A. Kumar, W. Wang, C. Long, M. Chelliah, Y. Xue, B. Huang, J.-K. Schemm, W. Ebisuzaki, R. Lin, P. Xie, M. Chen, S. Zhou, W. Higgins, C.-Z. Zou, Q. Liu, Y. Chen, Y. Han, L. Cucurull, R. W. Reynolds, G. Rutledge, and M. Goldberg, "The NCEP climate forecast system reanalysis," Bull. Amer. Meteorol. Soc., vol. 91, no. 8, pp. 1015–1058, Aug. 2010, doi: 10.1175/2010BAMS3001.1.

[3] S. Saha, S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H.-Y. Chuang, M. Iredell, M. Ek, J. Meng, R. Yang, M. P. Mendez, H. Van Den Dool, Q. Zhang, W. Wang, M. Chen, and E. Becker, "The NCEP Climate Forecast System version 2," J. Climate, vol. 27, no. 6, pp. 2185–2208, Mar. 2014, doi: 10.1175/JCLI-D-12-00823.1.

[4] K. E. Trenberth and J. T. Fasullo, "An apparent hiatus in global warming?," Earth's Future, vol. 1, no. 1, pp. 19–32, Mar. 2013, doi: 10.1002/2013EF000165.

[5] A. Aieb, K. Madani, M. Scarpa, B. Bonaccorso, and K. Lefsih, "A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria," Heliyon, vol. 5, no. 2, Art. no. e01247, Feb. 2019, doi: 10.1016/j.heliyon.2019.e01247.

[6] J. N. Valencia Gonzalez, R. A. Ramírez, M. A. V. Peña, and A. Quevedo Nolasco, "Relleno de datos diarios faltantes en registros de series climatológicas temporales," Rev. Mex. Cienc. Agríc., vol. 13, no. 4, pp. 617–629, Aug. 2022, doi: 10.29312/remexca.v13i4.2514.

[7] D. E. Booth, "Analysis of incomplete multivariate data," Technometrics, vol. 42, no. 2, pp. 213–214, May 2000, doi: 10.1080/00401706.2000.10486013.

[8] S. Ghosh, "Statistical analysis with missing data," Technometrics, vol. 30, no. 4, pp. 455–455, Nov. 1988, doi: 10.1080/00401706.1988.10488446.

[9] K. E. Ukhurebor, S. O. Azi, U. O. Aigbe, R. B. Onyancha, and J. O. Emegha, "Analyzing the uncertainties between reanalysis meteorological data and ground measured meteorological data," Measurement, vol. 165, Art. no. 108110, Jan. 2020, doi: 10.1016/j.measurement.2020.108110.

[10] J. L. Schafer, Analysis of Incomplete Multivariate Data. New York, NY, USA: Chapman & Hall/CRC, 1997. doi: 10.1201/9780367803025.

[11] C. Chatfield, "Prediction intervals for time-series forecasting," in Principles of Forecasting, J. S. Armstrong, Ed. Boston, MA, USA: Springer, 2001, pp. 475–494. doi: 10.1007/978-0-306-47630-3_21.

[12] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," Procedia Comput. Sci., vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[13] M. Mera-Gaona, U. Neumann, R. Vargas-Canas, and D. M. López, "Correction: Evaluating the impact of multivariate imputation by MICE in feature selection," PLOS One, vol. 16, no. 12, Art. no. e0261739, Dec. 2021, doi: 10.1371/journal.pone.0261739.

[14] S. Jäger, A. Allhorn, and F. Biessmann, "A benchmark for data imputation methods," Front. Big Data, vol. 4, Art. no. 693674, 2021, doi: 10.3389/fdata.2021.693674.

[15] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," J. Stat. Softw., vol. 85, no. 11, pp. 2541–2552, Nov. 2012, doi: 10.1016/j.jss.2012.05.073.

[16] R. J. Longman, A. J. Newsman, T. W. Giambelluca, and M. Lucas, "Characterizing the uncertainty and assessing the value of gap-filled daily rainfall data in Hawaii," J. Appl. Meteorol. Climatol., vol. 59, no. 7, pp. 1261–1276, Jul. 2020, doi: 10.1175/JAMC-D-20-0007.1.

[17] C. I. Anderson and W. A. Gough, "Accounting for missing data in monthly temperature series: Testing rule-of-thumb omission of months with missing values," Int. J. Climatol., vol. 38, no. 13, pp. 4990–5002, Nov. 2018, doi: 10.1002/joc.5801.

Francisco García Barrera received the B. Eng. degree in computer science from Universidad Arturo Prat, Iquique, Chile, and the M.S. degree in computer management from the same institution. He is currently an Assistant Professor with the Faculty of Engineering and Architecture, Universidad Arturo Prat, Iquique, Chile. His research interests include time-series analysis, data imputation, the use of reanalysis datasets, and data mining for hydrometeorological applications, with emphasis on predictive and descriptive modeling to reveal latent patterns. Prof. García Barrera has published in indexed journals and international conferences.

David Contreras Aguilar received his B.Eng. in Computer Science from Universidad Católica del Norte, Chile. He obtained a Master's degree in Information Technologies from Universidad Técnica Federico Santa María and a Ph.D. in Engineering from the Universidad de Barcelona, Spain. He is currently an Assistant Professor at the School of Engineering of Universitat Ramon Llull. He is a member of the Smart Society Research Group at the Universitat Ramon Llull, the Centre de Llenguatge i Computació (CLiC) and the WAI Research Group at the Universidad de Barcelona. His research interests include artificial intelligence systems, recommender systems, and fairness in recommendations. He has authored publications in peer-reviewed scientific journals and international conferences, and actively serves as a reviewer for leading international academic journals.

Héctor Aldea Navarro received his B.Sc. in Computer and Informatics Engineering from Universidad Arturo Prat, Iquique, Chile, in 2024. He is currently working as a Full-Stack Developer, contributing to the development of web applications for the mining industry and mutual safety services. His work includes the implementation of OCR-based APIs and the integration of advanced health monitoring solutions. His areas of interest include data processing, software development, and applied artificial intelligence.

Pablo Cárcamo Zuñiga received his degree in Civil Engineering in Computer Science from the Universidad Arturo Prat, Chile. He currently works as a Learning Analytics Specialist at the Center for Innovation and Teaching Development at the Universidad Católica de la Santísima Concepción. He has led the design and implementation of Power BI dashboards for analyzing LMS (Moodle) usage, faculty development, gender equity, and community engagement. His areas of interest include educational data analysis, institutional indicators, digital transformation in higher education, and business intelligence applied to academic decision-making.

Mauricio Oyarzún Silva received the degree of Civil Engineer in Computer Science and Informatics and the Ph.D. in Engineering Sciences with a specialization in Computer Science from the University of Santiago de Chile. He is currently a full-time faculty member at the Universidad Arturo Prat, Iquique, Chile. His research interests include information retrieval, compressed data structures, discrete-event simulation, and applications of artificial intelligence in engineering.

Alonso Inostrosa-Psijas received the Ph.D. degree from Universidad de Santiago de Chile, Chile. He is an Associate Professor at the School of Informatics Engineering at Universidad de Valparaíso, Chile. His research interests are discrete-event and parallel/distributed simulation. He can be reached at alonso.inostrosa@uv.cl.

Gabriel Icarte Ahumada received his B.Eng.in Computer Science from the Universidad Católica del Norte, Chile. He obtained an Master degree in Information Technologies from the Universidad Técnica Federico Santa María and a Doctor in Engineering degree from the Universidad de Bremen, Germany. He is currently an assistant Professor in the Faculty of Engineering and Architecture at the Universidad Arturo Prat. His research interests include multi agent systems, reinforcement learning, intelligent scheduling, and real time logistics applications in mining and transportation. He has published in indexed journals and international conferences.

Francisco Moreno Herrera is a Bachelor in Computer Science and Ph.D. in Engineering Sciences, teaches at the Mathematics and Computer Science Department in the Universidad de Santiago de Chile, Chile. He currently works in applied statistics and information theory.