

Trade-Offs in Statistical and Machine-Learning-Based Imputation for Algal Time-Series: A Multi-Criteria Evaluation on Optical Density Data Without Ground Truth

Erickson James¹, Ervin Gubin MOUNG^{2,*}, Mohd Shahazureen Ikwana Bin Abdul Rahman³,
Ahmad Asnawi Bin Mus⁴, Valentino Liaw⁵, Ali Farzamnia⁶, Lucky Goh Poh Wah⁷

^{1,2,3,5}*Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia*

^{4,7}*Faculty of Science and Technology, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia*

⁶*School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, United Kingdom*

(Received: April 20, 2026; Revised: June 20, 2026; Accepted: September 20, 2026; Available online: October 14, 2026)

Abstract

Missing values are pervasive in algal cultivation monitoring, where instrument failure and transmission loss remove whole runs of consecutive observations, and the repair method chosen silently shapes every downstream analysis. This study asks which repair method should be used, and on what evidence, when no complete reference series exists. Twelve methods are compared on optical density and ten covariates recorded across 30 open-pond production cycles at five geographically distinct sites: four interpolation schemes, a propagation scheme, two nearest-neighbor schemes, three chained-equation variants, a random-forest ensemble and a mechanistic growth-curve fit. The principal contribution is a validation framework organized in two clearly separate tiers. The first tier needs no ground truth and applies to the naturally missing values, comparing central tendency, dispersion, distributional shape, autocorrelation structure and coverage between observed and repaired series. The second tier withholds values that were in fact recorded and rebuilds them, so reconstruction error can be measured directly; it is replicated over ten random seeds, three masking proportions and four contiguous outage lengths, giving 25,200 paired evaluations in which every method sees an identical pattern of withhold cells. Key findings found that no method dominates. Ensemble and chained-equation learners rebuild individual values most accurately, with a normalized error of 0.1927 against 0.2645 for linear interpolation, but they compress dispersion and distort distributional shape, whereas shape-preserving interpolation and propagation protect those properties and rebuild less accurately. Gap structure matters: interpolation loses accuracy when gaps are contiguous rather than scattered, while multivariate learners gain. Regularizing the chained equations removes a convergence failure that left one pond unrepaired. An aggregated decision analysis turns these trade-offs into a weight-dependent selection rule, giving practitioners an auditable basis for matching method to analytical objective instead of a single universal default.

Keywords: Autocorrelation, Imputation, Missing Data, Multi-Criteria Decision Making, Optical Density

1. Introduction

In real-world scenarios, the failure of instrument functionality, inconsistent sampling, and transmission errors cause environmental and ecological time-series datasets to frequently suffer from missing values, which can distort statistical inference and reduce the reliability of analyses [1]. In ecological monitoring studies, ignoring or incorrectly handling missing data can produce biased outcomes, weaken statistical power, and alter ecological interpretation [2]. For this reason, data gaps are often ignored or simply removed from analysis, potentially discarding valuable information and introducing systematic bias.

There is substantial research on missing-data imputation for time-series across environmental sciences and related domains [3], [4]. Traditional techniques such as linear interpolation, spline methods, and moving averages have long been applied because of their simplicity and interpretability [4]. Statistical methods such as Kalman filtering and expectation–maximization approaches have been developed to better account for temporal dependence [5]. With the recent advancement of machine-learning methods, approaches such as K-Nearest Neighbours (KNN) and Multiple

*Corresponding author: Ervin Gubin MOUNG (ervin@ums.edu.my)

DOI: <https://doi.org/10.47738/jads.v7i4.1539>

This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights

Imputation by Chained Equations (MICE) have shown improved performance on environmental time series with large continuous missing segments [6], and recurrent and attention-based formulations have been developed for multivariate settings [7]. MissForest [8] and related random-forest approaches have also demonstrated strong performance on environmental data-imputation tasks [9]. Comparative evaluations in aquatic production monitoring report the same ordering, machine-learning imputers reconstructing flow-through aquaculture records more accurately than classical interpolation [10].

The use of suboptimal methods that smooth out critical biological spikes or crashes leads to inaccurate predictions and flawed environmental-management decisions [11]; this risk is acute for algal-bloom forecasting, where data-driven forecasting models are increasingly deployed [12]. The lack of comparative analysis between missing-data-handling methods in algae growth, especially in incomplete datasets, is concerning [13]. Current literature often applies common imputation techniques without evaluating their performance against the unique nonlinear growth phases of algae. Algal biomass production follows distinct lag, exponential, and stationary phases, each with different autocorrelation structures that imputation methods must preserve [14].

Moreover, utilizing an imputation method that ignores temporal structure can cause the dataset to lose its "memory", producing biased results that fail to account for the cumulative effect of environmental stressors [15], [7]. Because algal growth is governed by complex cross-variable interactions and strong autocorrelations, missing-data-handling methods must preserve these temporal dependencies [15]. Comparative environmental-modelling studies have shown that different methods offer different strengths in capturing nonlinear relationships and sequential structure [4], further supporting the importance of selecting imputation methods according to the analytical objective and the temporal characteristics of the data.

In recent years, no study has yet evaluated imputation methods for algal optical-density time-series under a multi-criteria framework that assesses statistical fidelity, distributional similarity and temporal-dependency preservation together. Existing comparisons are either domain-agnostic [4], [7] or rely on synthetic missingness in which reconstruction accuracy can be measured against a known reference [3], whereas real cultivation datasets have no complete reference sequence, so accuracy there can only be probed indirectly, by withholding values that were in fact observed. Whether a ranking obtained in that way survives realistic burst missingness, repeated masking realizations and a change of cultivation site is equally untested. These two gaps, the absence of a validated ground-truth-free evaluation methodology and the absence of evidence on its robustness, constitute the problem addressed by the present study.

This paper contributes both a comparative evaluation and the validation framework that supports it. Twelve methods spanning interpolation-based, propagation-based, neighbor-based, chained-equation, ensemble tree-based and mechanistic families are evaluated on OD680 and ten covariates from 30 production ponds at all five ATP3 sites. Validation is organized in two explicitly separated tiers, a ground-truth-free assessment of the naturally missing values and a simulated-missingness experiment reproducing both scattered dropout and contiguous sensor outages over ten seeds and three masking fractions, with every metric and aggregation rule defined in closed form and the stability of the resulting ranking quantified. The trade-offs this exposes between mean preservation, variance fidelity, distributional similarity, temporal-structure preservation and point-level accuracy are aggregated into a weight-dependent selection guideline, which turns the comparison into an auditable basis for choosing a method.

2. Literature Review

2.1. Statistical Imputation Methods

Statistical imputation techniques form the foundation of missing-data handling in time-series analysis. Linear interpolation, which estimates missing values by connecting observed data points with straight-line segments, remains widely used because of its computational efficiency and interpretability [4]. Spline-based methods extend this approach by fitting piecewise polynomials that capture local curvature in the data, with cubic splines being the most common variant [9]. This led to an unconstrained s-spline that can oscillate freely between and beyond knots. Thus, shape-preserving alternatives such as the piecewise cubic Hermite interpolating polynomial (PCHIP) trade the second-derivative continuity of the natural cubic spline for monotonicity on monotone data, which removes the overshoot that

unconstrained splines exhibit near series boundaries. Global polynomial fitting offers a parametric alternative by modelling the entire time series as a single polynomial function. These interpolation-based methods operate exclusively on the temporal structure of the target variable and are well-suited to univariate time series with smoothly varying patterns.

Propagation-based methods, such as Last Observation Carried Forward (LOCF) and its variants, fill missing values with the most recent observation. Despite their simplicity, these methods can introduce substantial bias in temporally autocorrelated data by creating artificial constant segments that distort the underlying dynamics [4]. In environmental time series, LOCF and related fixed-fill schemes have been shown to perform poorly when missing gaps exceed short durations [16], [17].

2.2. Machine-Learning Imputation Methods

Multiple Imputation by Chained Equations (MICE) is a widely adopted multivariate imputation approach that models each incomplete variable as a function of all other variables through an iterative regression procedure [18]. MICE has demonstrated strong performance on environmental datasets, particularly in preserving overall mean structure [3]. However, its regression-based mechanism tends to compress variance by drawing imputed values toward conditional means, a limitation noted across multiple domains [18].

KNN imputation identifies similar observations in the feature space and aggregates their values to estimate missing entries. In environmental applications, KNN has shown robust performance across various missingness mechanisms and data types [19]. Belachsen and Broday [19] demonstrated that a weighted KNN variant with lagged inputs outperformed iterative tree-based methods for short missing gaps in PM2.5 monitoring data, and Umar and Gray [18] found KNN to be among the best methods for multivariate water-level data, particularly under missing-not-at-random mechanisms. Related scattered-data approximation techniques, including Gaussian process regression and support vector regression, have likewise been compared for reconstructing irregularly sampled environmental data [20].

Random-forest-based methods, particularly MissForest [8], have gained traction in environmental imputation because of their ability to handle mixed-type data and complex nonlinear relationships. Zhou et al. [9] demonstrated that MissForest effectively fills daily streamflow gaps, though performance degraded under strong human-activity influences. Diouf et al. [21] identified MissForest as the best method for temperature and humidity imputation across multiple missingness levels in West African meteorological data.

Two refinements of the chained-equation scheme are directly relevant to short environmental series. First, the conditional models can be regularized: a ridge or lasso backbone keeps the iterations well posed when a cycle contains few complete rows and strongly collinear predictors, a regime in which unregularized iterative regression is rank-deficient and may fail to return any value. Second, imputations can be drawn from the conditional predictive distribution rather than set to the conditional mean; adding a residual draw at each iteration is the standard remedy for the variance compression that otherwise accompanies regression-based imputation. A separate strand of work embeds domain knowledge directly in the imputation model: because microalgal batch cultivation follows a characteristic lag–exponential–stationary trajectory, a parameter-bounded logistic curve or the Baranyi–Roberts model can be fitted to the observed points and the gaps read from the fitted trajectory, which constrains the reconstruction to behavior that is physiologically admissible, being non-negative, bounded and monotone within a growth phase, at the price of a functional assumption that does not hold for variables which are not growth-driven, such as temperature or pH.

2.3. Validation Approaches and Research Gap

Validation of imputation methods is typically conducted through direct accuracy metrics (RMSE, MAE) on synthetic missingness scenarios where ground truth is available [3], [7]. However, many real-world environmental datasets lack complete reference sequences, necessitating indirect validation approaches. Distributional tests such as the Kolmogorov–Smirnov test, preservation of descriptive statistics, and autocorrelation analysis offer ground-truth-free alternatives. Despite the availability of these validation tools, no prior study has applied them systematically to compare imputation methods on real-world algal cultivation data, nor tested whether the resulting ranking is stable across masking realizations, gap structures and cultivation sites. Alejo-Sanchez et al. [22] recently reviewed imputation methods for climate time series and emphasized the need for domain-specific comparative studies with realistic

missingness patterns. The present study addresses this gap with a multi-criteria comparison of twelve imputation methods on algal OD680 time-series from 30 open-pond cultivation cycles at five geographical sites.

3. Methodology

The research design follows four sequential phases, namely data preparation, imputation, validation and synthesis, as summarized in [figure 1](#), with validation comprising both the ground-truth-free tier and the supplementary simulated-missingness tier defined in Section 3.4.1.

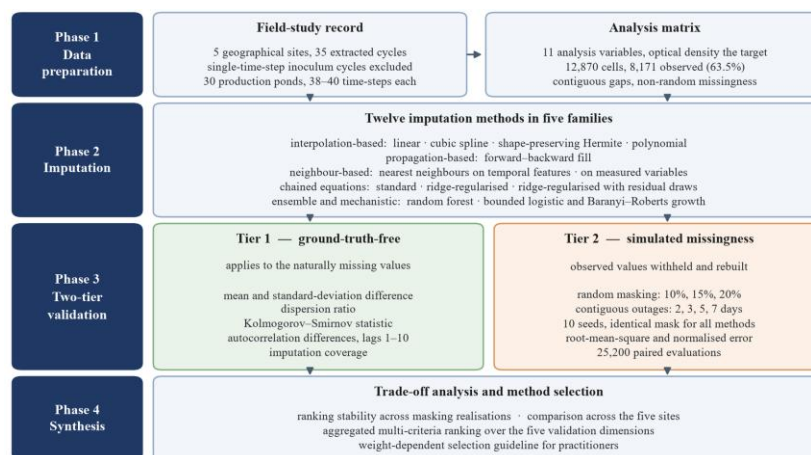


Figure 1. Research design and flow of the two-tier multi-criteria imputation evaluation, from cycle extraction through the twelve imputation methods and the two validation tiers to the aggregated method-selection guideline

3.1. Dataset Description

This study uses a publicly available dataset from the Algae Testbed Public-Private Partnership (ATP3), specifically the Unified Field Study 7 experiment [14]. The dataset comprises operational data from algae cultivation in open ponds conducted at five geographical sites: Arizona (ASU), California (CP), Florida (FA), Georgia (GT), and Hawaii (CELL). The present analysis uses all five sites. After excluding the five single-time-step inoculum cycles, 30 production ponds, six per site, enter the evaluation, each spanning 38–40 time-steps over a cultivation cycle.

The dataset includes 14 core numeric variables: optical density at 680 nm (OD680) and 750 nm (OD750), dry weight (DW.g.L), ash-free dry weight (AFDW.g.L), ash content, depth (cm), temperature (Temp.C), pH, salinity, nitrate (NO3.mg.L), phosphate (P.mg.L), evaporation rate (cm/day), and two derived variables (NO3.demand.g.N.g.AFDW and mg.AFDW_mol.PAR.input). OD680 serves as the primary variable of interest as it directly measures algal biomass concentration. The original dataset of 1,567 rows was reduced to 1,541 rows after removing 26 all-missing rows. Imputation coverage (Section 4.1) is reported over all 14 variables and all 35 extracted cycles, whereas the simulated-missingness grid is run on the 11 analysis variables retained by the preparation stage for every site, across the 30 production ponds. That grid comprises 1,170 pond–time-step rows and 12,870 variable cells, of which 8,171 (63.5%) are observed. [Table 1](#) presents the dataset statistics by site.

Table 1. Dataset statistics for the 30 ATP3 production ponds, by site

Site	Ponds	Time-steps	Variable cells	Observed	Missing	Missing %
Arizona (ASU)	6	40	2,640	2,032	608	23.03
Hawaii (CELL)	6	38	2,508	1,393	1,115	44.46
California (CP)	6	39	2,574	1,444	1,130	43.90
Florida (FA)	6	40	2,640	1,851	789	29.89
Georgia (GT)	6	38	2,508	1,451	1,057	42.15
All sites	30	38–40	12,870	8,171	4,699	36.51

Missingness across the 11 analysis variables ranges from 23.03% at the Arizona site to 44.46% at the Hawaii site, and OD680 itself is missing for 277 of 1,170 pond–time-steps (23.68%), leaving 26 to 36 observed OD680 points per pond. Missing values arise predominantly from instrument failure and transmission errors and therefore form contiguous gaps

rather than isolated dropout, a structure consistent with a non-MCAR (Missing Completely at Random) mechanism; a definitive MAR (Missing at Random) or MNAR (Missing Not at Random) classification would require formal diagnostics such as Little's test, which lie outside the scope of this study. The evaluation protocol of Section 3.4.4 does not depend on the label, since it reports performance separately under scattered random masking and under contiguous outages.

3.2. Data Preprocessing

The preprocessing pipeline began with the extraction of production cycles from the SiteID–PondID combination, yielding 35 distinct cycles across the five sites: six production cycles plus one inoculum cycle per site. Temporal alignment was performed using the Duration days column as the temporal index. The five inoculum cycles contain a single temporal point each and were excluded as insufficient for comparative analysis, leaving 30 production cycles. Duplicate rows were checked (none found) and all-missing rows were removed. Outlier detection was performed per numeric variable using an IQR-based approach with a threshold of three standard deviations, identifying 0–12 outliers per variable, with no variable exceeding 1.5% outlier prevalence.

3.3. Imputation Methods

Twelve imputation methods spanning five methodological categories were evaluated, all of them implemented within the same cycle-wise framework so that they receive identical inputs and are scored on identical masks. Positions a method cannot fill are left missing and reported as coverage rather than filled silently.

Four interpolation-based methods estimate missing values from the temporal structure of the target variable alone. Linear interpolation joins successive observed points with straight-line segments and is included as the simplest baseline in common use, while cubic spline interpolation fits a piecewise cubic through every observed point under a not-a-knot boundary condition, which follows local curvature but is free to oscillate near cycle boundaries. PCHIP, a shape-preserving piecewise cubic Hermite interpolant, is included precisely to test that weakness, being monotone wherever the data are monotone, unable to overshoot between knots, and clipped to physically valid non-negative optical densities. Polynomial fitting adds a single global cubic, degree three being the lowest order able to represent an inflection point together with the lag, exponential and early-stationary phases, a choice tested empirically in Section 4.2.

Forward–backward fill represents the propagation family, carrying the most recent observation forward and resolving any remaining gap from the next available value, and it is included because it is the default in many monitoring pipelines despite the artificial constant segments it creates. KNN represents the neighbor-based family in its conventional temporal form, regressing on the day index and its square and cube with $k = 3$ and inverse-distance weighting, so that neighborhood is defined by position within the growth cycle.

The three chained-equation variants treat each incomplete variable as a regression target predicted by the remaining variables of the same cycle, iterating at most 15 times to a tolerance of 10^{-3} . MICE is the standard configuration and serves as the multivariate baseline, but it returns nothing for a cycle containing no complete predictor rows. MICE-Ridge removes that failure mode by fitting the equations once over the whole cycle frame and completing any cell they cannot fill by linear interpolation; the regularization itself is not the repair, since the conditional model is a Bayesian ridge by default. MICE-Ridge stochastic draws each imputed value from the conditional predictive distribution instead of the conditional mean, by adding a residual draw scaled to the in-sample residual spread, to test whether the variance compression inherent in regression-based imputation can be relieved.

Three further methods exploit the measured feature space or biological knowledge directly. KNN multivariate defines neighborhood by environmental similarity rather than by proximity in time, with the neighborhood size selected per series by leave-one-out cross-validation over the values 1, 3, 5 and 7, and it tests whether the cross-variable structure carries information a time-only neighborhood discard. MissForest [8], an iterative random-forest imputer using 30 trees and at most 4 chained iterations, represents the ensemble family and imposes no functional form on the growth trajectory. The growth model fits a parameter-bounded logistic curve and a Baranyi–Roberts curve and reads the gaps from whichever attains the lower residual sum of squares, with the carrying capacity bounded to between one half and five times the observed maximum, so as to test whether embedding growth kinetics improves reconstruction; where

neither form converges it falls back to linear interpolation, which keeps its coverage comparable with the other eleven methods.

3.4. Validation Framework

3.4.1. Two-Tier Validation and Terminology

No complete reference sequence exists for the naturally missing observations, so imputation quality for those values can only be assessed indirectly, and validation is therefore organized in two tiers that are kept terminologically distinct throughout. The first tier is ground-truth-free and applies to the naturally missing values, comparing the statistical, distributional and temporal properties of the imputed series with those of the observed series. The second tier is a supplementary simulated-missingness experiment in which values that were in fact observed are withheld, imputed and compared with the withheld observations; because those values are genuine measurements the tier yields direct error metrics, but they are pseudo-ground-truth in that the missingness is simulated, and they describe reconstruction of observable cells rather than of the naturally missing ones. Wherever RMSE or NRMSE is reported it belongs to the second tier, and the term “ground-truth-free” is reserved for the first.

3.4.2. Validation Criteria

The ground-truth-free tier assesses imputation quality across four complementary dimensions, and coverage is recorded alongside them as an operational reliability measure. Descriptive-statistics preservation: The mean and standard deviation are computed for the original observed values and the corresponding imputed datasets. The absolute differences in mean and SD quantify how well each method preserves the central tendency and dispersion of the original distribution. Kolmogorov–Smirnov (KS) test: The two-sample KS test compares the empirical cumulative distribution functions of the original observed values and the imputed values. The KS statistic D measures the maximum vertical deviation between the two distributions, and the associated p-value indicates whether the difference is statistically significant at $\alpha = 0.05$. Autocorrelation function (ACF) analysis: The ACF is computed for both the original and imputed time series across lags 1–10. Two metrics are derived: the maximum absolute ACF difference across all lags (capturing worst-case temporal distortion) and the mean absolute ACF difference (capturing overall temporal-structure preservation). Smaller values indicate better preservation of temporal dependencies. Coverage: The proportion of originally missing cells for which a method returns a value is recorded per method, per cycle and per variable. Coverage is not an accuracy measure, but a method that silently returns nothing on part of the data cannot be compared with one that does not, and in an operational pipeline such a failure propagates as data loss.

3.4.3. Formal Definitions

This subsection defines the quantities reported in Section 4, together with the rules used to aggregate them to pond level and overall. Let $\hat{y}_{pvt}^m$ be the value produced by method m for variable v of pond p at time-step t, O_{pv} the set of time-steps at which that variable is observed, and $\Omega_{pv} \subseteq O_{pv}$ the subset withheld by the masking protocol. Equation (1) gives the reconstruction error at the withheld positions, Equation (2) normalizes it by the observed range of that variable in that pond so that variables measured on different scales are comparable, and Equation (3) aggregates it, first over the V_p variables for which it is defined and then over the P ponds and the S masking realizations.

$$RMSE_{mpv} = \sqrt{\left[(1 / |\Omega_{pv}|) \cdot \sum (\hat{y}_{pvt}^m - y_{pvt})^2 \right]}, \quad t \in \Omega_{pv} \quad (1)$$

$$NRMSE_{mpv} = RMSE_{mpv} / (\max y_{pvt} - \min y_{pvt}), \quad t \in O_{pv} \quad (2)$$

$$NRMSE_{mp} = (1 / V_p) \cdot \sum_v NRMSE_{mpv} \quad \text{and} \quad NRMSE_m = (1 / (P \cdot S)) \cdot \sum_p \sum_s NRMSE_{mp}^{(s)} \quad (3)$$

For the ground-truth-free tier, let y_p^{obs} be the observed values of pond p, $\hat{y}_p^m$ the corresponding completed series, and I_p the positions at which the target variable is missing. Equation (4) gives the absolute differences in the first two moments together with the dispersion ratio, for which unity denotes exact preservation, and Equation (5) the two-sample Kolmogorov–Smirnov statistic between the empirical distribution function F_p of the observed values and the function G_{mp} of the values inserted at I_p , evaluated at $\alpha = 0.05$. Equation (6) gives the worst-case and mean absolute distortion of the sample autocorrelation ρ_k over lags $k = 1$ to $K = 10$, computed once on the observed values and once on the completed series, and Equation (7) gives imputation coverage, the percentage of originally missing cells that a method fills.

$$\Delta\mu_{mp} = |\mu(y_{p}^{obs}) - \mu(\hat{y}_{p}^m)|, \quad \Delta\sigma_{mp} = |\sigma(y_{p}^{obs}) - \sigma(\hat{y}_{p}^m)|, \quad R\sigma_{mp} = \sigma(\hat{y}_{p}^m) / \sigma(y_{p}^{obs}) \quad (4)$$

$$D_{mp} = \sup_x |F_p(x) - G_{mp}(x)| \quad (5)$$

$$\Delta ACF_{mp}^{max} = \max_k |\rho_k(y_{p}^{obs}) - \rho_k(\hat{y}_{p}^m)| \quad \text{and} \quad \Delta ACF_{mp}^{mean} = (1 / K) \cdot \sum_k |\rho_k(y_{p}^{obs}) - \rho_k(\hat{y}_{p}^m)| \quad (6)$$

$$COV_m = 100 \cdot (\text{cells filled by method } m) / (\text{originally missing cells}) \quad (7)$$

3.4.4. Simulated-Missingness Protocols

Two masking schemes are applied to every production pond. Under random masking a fraction f of the observed cells of each variable is withheld uniformly at random, with f varied over 10%, 15%, 20% and the draw repeated for 10 seeds, giving 30 realizations per pond. Under block masking one contiguous outage of L consecutive sampling days is placed at a random start position in each variable, with $L \in \{2, 3, 5, 7\}$ days and the same 10 seeds, giving 40 realizations per pond; only cells that were actually observed are scored, and a placement that would withhold every observation of a variable is rejected and redrawn. The block scheme is the operationally realistic one, since sensor failure, biofouling and telemetry outages in open-pond monitoring produce multi-day gaps rather than isolated dropouts. Within each realization all twelve methods receive the same mask, so that every comparison between two methods is like-for-like. A single-realization protocol, using one seed and 15% of observed values withheld under a no-two-consecutive constraint on the six Arizona ponds, is reported alongside the grid as a reference configuration, with per-pond and per-variable values in Supplementary Tables S8 and S9.

3.5. Experimental Setup

All experiments were conducted on the 30 production ponds across the five sites, with OD680 as the primary analysis variable and the remaining 10 analysis variables imputed and scored alongside it. Imputation was performed independently within each production cycle to preserve the biological boundaries between cultivation batches, under identical preprocessing and temporal-alignment conditions for every method. The simulated-missingness grid comprises 30 ponds by 10 seeds by 3 masking fractions by 12 methods, giving 10,800 random-masking records, and 30 ponds by 10 seeds by 4 outage lengths by 12 methods, giving 14,400 block-masking records, for 25,200 paired evaluations in total. Results are checkpointed per pond-configuration, and all implementation was carried out in Python using the pandas, scipy, numpy and scikit-learn libraries with a fixed master seed of 42.

4. Results and Discussion

4.1. Imputation Coverage and Operational Reliability

Table 2 reports imputation coverage over all 14 variables and all 35 extracted cycles. Forward-backward fill imputes the highest proportion of missing values (90.11%), since propagation requires only one observed neighbor in either direction, while baseline MICE imputes the fewest (82.98%) leaving 1,168 gaps. This could be caused by a production cycle containing no complete rows that gives the chained equations no usable predictors. The remaining ten methods lie between 85.63% and 86.16%.

The consequence is visible at pond level. Under the single-realization protocol baseline MICE returns no value for any of the 50 withheld cells of ASU_P1 whereas every other method recovers all of them. Fitting the chained equations once over the whole cycle frame, adding a ridge backbone and falling back to linear interpolation for any unresolved cell removes the failure, so that MICE-Ridge recovers all 50 cells at an NRMSE of 0.1577 and whole-dataset coverage rises from 82.98% to 86.16%. Over the 30 production ponds and 11 analysis variables of the evaluation grid every method reaches 100% coverage, which localizes the failure to cycles in which the target variable is the only one observed.

Because MICE produces no output for ASU_P1, an average over the Arizona ponds would rest on five ponds for MICE and six for every other method. Restricting all methods to ASU_P2 to ASU_P6 changes each mean NRMSE by at most 0.0050 and leaves MICE best under both scopes, at 0.1723 over six ponds and 0.1740 over five, and recomputing every method on the 638 pond-configuration cells for which MICE returned a value gives a dropout bias between -0.0057 and 0.0000 for all methods except cubic spline. The advantage is therefore not a sampling artefact, although it remains

conditional on convergence, and MICE-Ridge trades accuracy for guaranteed output at 0.2308 against 0.1723. Supplementary Table S3 gives full experimental results.

Table 2. Imputation coverage across all 14 variables and 35 cultivation cycles

Method	Points imputed	Imputation rate (%)	Remaining missing	Arizona ponds returning no output
Linear interpolation	5,913	86.16	950	0 of 6
Polynomial	5,877	85.63	986	0 of 6
Cubic spline	5,877	85.63	986	0 of 6
MICE	5,695	82.98	1,168	1 of 6
Forward-backward fill	6,184	90.11	679	0 of 6
KNN	5,913	86.16	950	0 of 6
PCHIP	5,913	86.16	950	0 of 6
MICE-Ridge	5,913	86.16	950	0 of 6
MICE-Ridge stochastic	5,913	86.16	950	0 of 6
KNN multivariate	5,877	85.63	986	0 of 6
MissForest	5,913	86.16	950	0 of 6
Growth model	5,913	86.16	950	0 of 6

4.2. Empirical Selection of the Polynomial Degree

The degree of the global polynomial is selected by leave-one-out cross-validation rather than on the conceptual grounds that a cubic is the lowest order able to represent an inflection point together with the lag, exponential and early-stationary phases. For every pond-variable series the polynomial is refitted with one observation held out at a time, for degrees one to five, and the out-of-sample NRMSE is averaged over the held-out points. For OD680 the mean across the 30 ponds falls from 0.2584 at degree one to 0.2027 at degree three and rises at degrees four (0.2130) and five (0.2285), and over all 325 scored series the ordering is the same, at 0.2676, 0.2677, 0.2596, 0.2949 and 0.3752, so degree three is both the empirical optimum and the parsimonious choice and is retained.

The optimum is not uniform across variables, degree three winning for the 5 biomass-related variables (AFDW.g.L, Ash, DW.g.L, OD680, OD750) and outright on 104 of the 325 scored series, degree one for the monotone or near-linear variables (Depth.cm, NO3.demand.g.N.g.AFDW, Temp.C, mg.AFDW_mol.PAR.input) and on 95 series, degree two for pH, and degree four for NO3.mg.L, which wins 45 series. Supplementary Table S1 gives the per-variable values.

4.3. Ground-Truth-Free Preservation of Statistical, Distributional and Temporal Structure

Table 3 summarizes the four ground-truth-free criteria for OD680 averaged over the 30 production ponds. MICE achieves the lowest mean difference (0.0030), followed by the growth model (0.0032), MICE-Ridge (0.0036) and MICE-Ridge stochastic (0.0038), since conditioning on the other measured variables estimates the conditional mean directly. The same mechanism compresses dispersion, MICE recording the largest SD difference (0.0196) at a dispersion ratio of 0.8954, with KNN multivariate (0.0193, ratio 0.8756) and MissForest (0.0193, ratio 0.8777) close behind. The interpolation and propagation methods invert the pattern, PCHIP having the smallest SD difference (0.0056) at a ratio of 1.0009, followed by linear interpolation (0.0057) and forward-backward fill (0.0060), at the cost of mean differences two to three times larger. Cubic spline alone inflates dispersion (ratio 1.1050), which Section 4.5 traces to boundary overshoot.

The Kolmogorov-Smirnov column carries two qualifications. A non-significant result indicates insufficient evidence to reject equality rather than equivalence, and with 26 to 36 observed points per pond the test has limited power, so the magnitude of D rather than the verdict is informative, and the verdict is in any case not uniform across the 30 ponds, non-rejection rates ranging from 33.3% for MICE and MissForest to 100% for MICE-Ridge stochastic. On magnitude, forward-backward fill matches the observed distribution most closely (0.2868), followed by KNN (0.3039), polynomial fitting (0.3138) and linear interpolation (0.3231), while MissForest (0.5506), MICE (0.5265), KNN multivariate (0.5173) and MICE-Ridge (0.5067) lie furthest away, which is the distributional counterpart of their variance compression.

Drawing imputations from the conditional predictive distribution rather than the conditional mean addresses both symptoms, since adding a residual draw to the ridge-regularized chained equations lowers the SD difference from

0.0187 to 0.0163, moves the dispersion ratio from 0.8834 to 0.8998, lowers the KS statistic from 0.5067 to 0.3454 and raises the non-rejection rate from 46.7% to 100%, at a cost in point accuracy reported in Section 4.4. Autocorrelation separates the methods least: the growth model preserves the profile best on the mean criterion (0.1193), ahead of KNN multivariate (0.1247), MICE (0.1248) and MissForest (0.1256), with cubic spline worst (0.1361), but the whole spread of 0.0168 is about 13.1% of its own average against a spread of 0.2638 on the KS statistic, so temporal preservation should not by itself drive method selection at this series length.

Table 3. Ground-truth-free validation criteria for OD680, averaged over 30 production ponds

Method	Δ mean	Δ SD	SD ratio	KS D	Non-rejection (%)	Max Δ ACF	Mean Δ ACF
Linear interpolation	0.0103	0.0057	0.9933	0.3231	96.7	0.2999	0.1296
Polynomial	0.0041	0.0104	0.9341	0.3138	96.7	0.2873	0.1288
Cubic spline	0.0182	0.0182	1.1050	0.3875	83.3	0.3361	0.1361
MICE	0.0030	0.0196	0.8954	0.5265	33.3	0.2775	0.1248
Forward-backward fill	0.0102	0.0060	0.9800	0.2868	96.7	0.3060	0.1286
KNN	0.0074	0.0071	0.9640	0.3039	96.7	0.2883	0.1309
PCHIP	0.0111	0.0056	1.0009	0.3287	93.3	0.3022	0.1291
MICE-Ridge	0.0036	0.0187	0.8834	0.5067	46.7	0.2830	0.1281
MICE-Ridge stochastic	0.0038	0.0163	0.8998	0.3454	100.0	0.2951	0.1286
KNN multivariate	0.0074	0.0193	0.8756	0.5173	40.0	0.2797	0.1247
MissForest	0.0078	0.0193	0.8777	0.5506	33.3	0.2868	0.1256
Growth model	0.0032	0.0155	0.9004	0.4969	43.3	0.2788	0.1193

4.4. Simulated-Missingness Accuracy under Random Masking

Table 4 reports mean NRMSE under random masking over 30 ponds, 10 seeds and 3 masking fractions, together with the distribution of per-configuration ranks. MissForest attains the lowest mean NRMSE at 0.1927 ± 0.0361 , marginally ahead of MICE at 0.1933 ± 0.0458 and clearly ahead of KNN multivariate at 0.2085 ± 0.0351 . Defining the neighborhood over the measured environmental variables and selecting k by cross-validation improves KNN from 0.2304 to 0.2085, a reduction of 9.5%, which shows that the cross-variable correlation structure carries information a time-only neighborhood discards. The growth model (0.2402) falls between polynomial fitting (0.2396) and linear interpolation (0.2645), so embedding growth kinetics is competitive with statistical multivariate learning without beating it at this series length. Cubic spline is an outlier at 1.0868 ± 1.7306 , whose worst configuration reaches 25.78.

The results also show that error grows monotonically with the masking fraction, from 0.1793 at 10% to 0.2033 at 20% for MissForest and from 0.2533 to 0.2749 for linear interpolation. However, the ordering is stable rather than an artefact of one draw: Kendall's coefficient of concordance across the 30 configurations is 0.955, no method has a rank standard deviation above 1.22, MissForest ranks first in 22 configurations and MICE in the remaining 8, and cubic spline ranks twelfth in all 30. Point accuracy is therefore led jointly by MICE and MissForest, which are separated by 0.0006 in mean NRMSE, an order of magnitude less than the gap to the third-placed method, with the leader depending on the masking protocol (Section 4.5).

Table 4. Mean NRMSE under random masking and stability of the resulting ranking across the 30 masking configurations

Method	Mean $\pm$ SD	10% masked	15% masked	20% masked	Mean rank	Rank SD	Times ranked 1st
Linear interpolation	0.2645 ± 0.0620	0.2533	0.2654	0.2749	7.97	0.89	0/30
Polynomial	0.2396 ± 0.0844	0.2251	0.2435	0.2501	5.37	1.22	0/30
Cubic spline	1.0868 ± 1.7306	0.9376	1.0919	1.2311	12.00	0.00	0/30
MICE	0.1933 ± 0.0458	0.1802	0.2009	0.2069	1.83	0.59	8/30
Forward-backward fill	0.2740 ± 0.0475	0.2626	0.2757	0.2835	9.10	0.92	0/30
KNN	0.2304 ± 0.0399	0.2186	0.2348	0.2378	4.30	0.60	0/30
PCHIP	0.2845 ± 0.0611	0.2716	0.2874	0.2943	10.43	0.68	0/30
MICE-Ridge	0.2531 ± 0.0486	0.2364	0.2551	0.2680	6.93	0.83	0/30
MICE-Ridge stochastic	0.2796 ± 0.0493	0.2637	0.2817	0.2934	10.10	0.88	0/30
KNN multivariate	0.2085 ± 0.0351	0.1926	0.2128	0.2199	2.90	0.31	0/30
MissForest	0.1927 ± 0.0361	0.1793	0.1953	0.2033	1.27	0.45	22/30

Growth model	0.2402 ± 0.0450	0.2304	0.2415	0.2487	5.80	0.81	0/30
--------------	-----------------	--------	--------	--------	------	------	------

4.5. Simulated-Missingness Accuracy under Contiguous Sensor Outages

Random masking leaves an observed neighbor on each side of almost every gap, which is the configuration interpolation is built for, whereas sensor failure, biofouling and telemetry outages produce multi-day gaps. Table 5 repeats the experiment with one contiguous outage of 2, 3, 5 or 7 consecutive sampling days per variable over the same 10 seeds and reports the penalty relative to random masking. The penalty falls on the interpolation methods, linear interpolation degrading by 2.41% and polynomial fitting by 3.03% while MICE improves by 4.77%, KNN multivariate by 3.73% and MissForest by 3.38%, because a contiguous outage leaves the cross-variable predictors intact while removing exactly the temporal neighbors interpolation depends on. MICE consequently overtakes MissForest under burst missingness, at 0.1841 against 0.1861, and the distortion is one of magnitude rather than order, since only 6 of the twelve methods change rank between protocols and none by more than 1 position.

Degradation with outage length is severe for the unconstrained spline and negligible for its shape-preserving replacement. Cubic spline rises from 0.6999 at a two-day outage to 4.4916 at a seven-day outage, an overall penalty of 93.00%, because a natural cubic spline extrapolating across a week of missing days oscillates freely and produces unphysical optical densities, whereas PCHIP, which is monotone on monotone data, cannot overshoot between knots and is clipped to non-negative values, rises only from 0.2444 to 0.3194 and carries a small negative penalty of -0.65%. Replacing the spline with a shape-preserving interpolant therefore removes the boundary instability at a modest cost in average accuracy relative to linear interpolation, at 0.2826 against 0.2709, and the pattern holds at every site (Supplementary Table S2).

Table 5. Mean NRMSE under contiguous outages of 2, 3, 5 and 7 sampling days, and the penalty relative to random masking

Method	2-day	3-day	5-day	7-day	All outages	Penalty vs random (%)	Rank shift
Linear interpolation	0.2292	0.2487	0.2862	0.3195	0.2709	+2.41	-1
Polynomial	0.2113	0.2242	0.2589	0.2930	0.2468	+3.03	-1
Cubic spline	0.6999	0.9526	2.2463	4.4916	2.0976	+93.00	0
MICE	0.1684	0.1777	0.1967	0.2068	0.1841	-4.77	+1
Forward-backward fill	0.2458	0.2542	0.2843	0.2981	0.2706	-1.22	+1
KNN	0.2062	0.2161	0.2367	0.2503	0.2273	-1.34	0
PCHIP	0.2444	0.2648	0.3019	0.3194	0.2826	-0.65	0
MICE-Ridge	0.2212	0.2342	0.2626	0.2793	0.2493	-1.51	0
MICE-Ridge stochastic	0.2487	0.2617	0.2872	0.3012	0.2747	-1.75	0
KNN multivariate	0.1810	0.1898	0.2096	0.2224	0.2007	-3.73	0
MissForest	0.1676	0.1752	0.1952	0.2066	0.1861	-3.38	-1
Growth model	0.2185	0.2269	0.2504	0.2659	0.2404	+0.09	+1

4.6. Cross-Site Generalization

Supplementary Table S4 reports mean NRMSE under random masking by site. Difficulty varies, reconstruction being easiest at Hawaii (0.2227) and hardest at California (0.2608) when averaged over the eleven non-degenerate methods, which broadly tracks the site missingness rates of Table 1, but the ordering of methods is conserved. MissForest is best at California, Florida and Georgia and MICE at Arizona and Hawaii, the two never differing by more than 0.0149 at any site, KNN multivariate is third at Arizona, California and Florida, second at Georgia and fifth at Hawaii, linear interpolation never ranks better than seventh, and cubic spline is last at all five. The trade-off structure therefore holds across five environmental regimes, which strengthens the recommendations without removing the caveat that all thirty cycles come from one field-study programme with a single collection period per site.

4.7. Multi-Criteria Decision Analysis

Sections 4.3 to 4.6 establish that the criteria disagree, and a multi-criteria decision analysis makes that disagreement operational. The decision matrix comprises the twelve methods and five cost-type criteria, namely mean difference, SD difference, KS statistic, mean autocorrelation difference and masked NRMSE under random masking (Supplementary Table S5), and is evaluated by the technique for order of preference by similarity to the ideal solution

under four weighting scenarios. Table 6 states the resulting selection guideline, figure 2 shows the closeness coefficients and Supplementary Table S6 provides full experimental results.

Table 6. Method-selection guideline derived from the TOPSIS analysis

Analytical objective	Criterion weighting	TOPSIS choice	C_i	Runner-up	Grey relational check
Equal weights	0.20 on each of the five criteria	Polynomial fitting	0.8819	KNN	differs – Forward–backward fill
Point-accuracy focus	0.50 NRMSE, 0.20 Δ mean, 0.10 each on Δ SD, KS, Δ ACF	Polynomial fitting	0.9392	Growth model	differs – MICE
Variance focus	0.35 Δ SD, 0.35 KS, 0.10 each on Δ mean, Δ ACF, NRMSE	KNN	0.8717	Forward–backward fill	differs – Forward–backward fill
Temporal focus	0.45 Δ ACF, 0.15 each on Δ SD, KS, NRMSE, 0.10 Δ mean	Polynomial fitting	0.8585	KNN	differs – Growth model

Under equal weights polynomial fitting attains the highest closeness coefficient (0.8819), ahead of KNN (0.8343) and MICE-Ridge stochastic (0.7756), while the methods that dominate raw point accuracy fall to seventh for MICE (0.7218) and eleventh for MissForest (0.6728) once their variance and distributional penalties are counted. Weighting point accuracy at 0.50 lifts the growth model to second (0.9164) and MICE to third (0.9149) but leaves polynomial fitting first (0.9392) because it is never worst on any criterion, weighting variance and distributional similarity at 0.35 each makes KNN the clear choice (0.8717) ahead of forward–backward fill (0.8395), and a temporal weighting returns polynomial fitting (0.8585), a scenario that should carry least weight given the small separation on that criterion.

Through two sensitivity checks, changes have been observed in these rankings. Cubic spline is a degenerate alternative whose masked NRMSE exceeds twice the median and compresses the remaining methods into a narrow band of the normalization. Removing it leaves polynomial fitting the winner under equal weights and under the temporal weighting, and KNN the winner under the variance weighting, but changes the point-accuracy winner to MICE, whose closeness coefficient is 0.7212 once cubic spline is excluded. A grey relational analysis on the same matrix agrees with the closeness ordering, at Spearman ρ between +0.755 and +0.972, yet selects a different winner in every scenario (Supplementary Table S7), so an aggregated index narrows the field rather than declaring one method optimal, and the rank spread across scenarios, from 0 for cubic spline to 8 for forward–backward fill, measures how weight-sensitive each recommendation is.

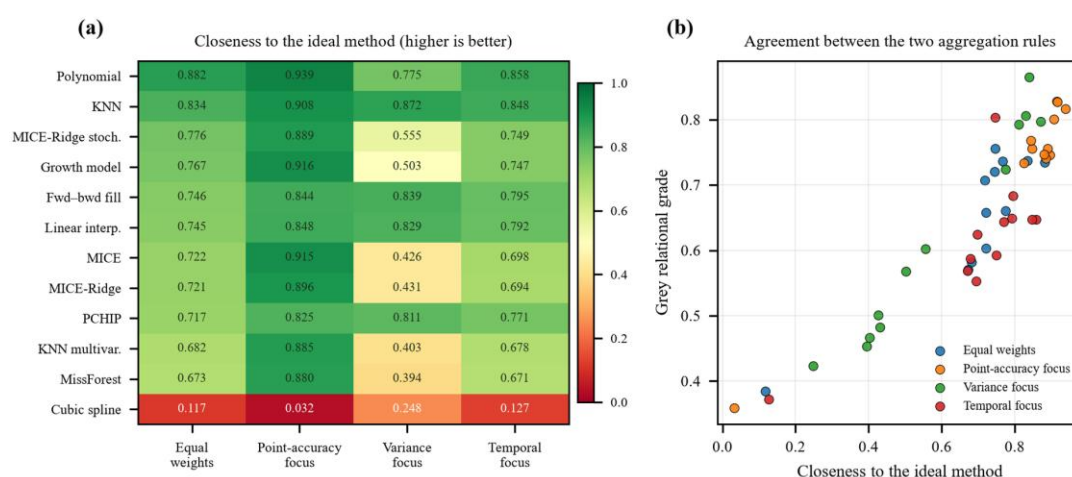


Figure 2. Aggregated multi-criteria ranking of the twelve methods. (a) Closeness to the ideal method under each of the four weighting scenarios. (b) Agreement between the closeness coefficient and the grey relational grade

4.8. Synthesis and Discussion

The twelve methods form a trade-off structure with three poles. The conditional-mean learners, namely MICE, MICE-Ridge, MissForest and KNN multivariate, dominate point reconstruction and mean preservation and are the only methods that improve under burst missingness, but they compress variance to dispersion ratios of 0.8756 to 0.8954 and produce the largest distributional deviations, at KS statistics of 0.5067 to 0.5506. The shape-preserving and propagation

methods, namely PCHIP, linear interpolation, forward-backward fill and KNN, occupy the opposite pole, preserving dispersion at ratios of 0.9640 to 1.0009 and distributional shape at KS statistics of 0.2868 to 0.3287 but giving up point accuracy, while polynomial fitting and the growth model sit between the two, which is why an equally weighted index selects them although neither wins any single criterion.

The study draws three consequences for practice. Variance preservation belongs to PCHIP, which has the smallest SD difference (0.0056) and the dispersion ratio closest to unity (1.0009), ahead of linear interpolation (0.0057), forward-backward fill (0.0060) and KNN (0.0071), so PCHIP is the recommendation wherever peak detection, extreme-event identification or bloom-onset flagging demands variance fidelity, with KNN preferred only when a multivariate reconstruction is also required and cubic spline disqualified by its behavior under outages despite adequate dispersion on scattered gaps. Point accuracy is shared between MICE and MissForest and is conditional on convergence, so only the ridge-regularized whole-frame variant with a deterministic fallback is safe to deploy unattended. Linear interpolation is not the general-purpose default it is often taken to be, ranking eighth of twelve under random masking, ninth under contiguous outages and sixth on the equally weighted index, and polynomial fitting and KNN are the better-balanced defaults within the conditions studied.

Method rankings are stable across ten masking realizations, at a coefficient of concordance of 0.955, and across three masking fractions, two gap structures and five sites, which is a demanding test of robustness, but all 30 cycles come from one dataset, one experimental programme and one measurement protocol, and each cycle spans only 38 to 40 sampling days, so the recommendations are scoped accordingly.

4.9. Limitations and Future Work

The study has found three limitations that bound the evidence. The evaluation covers 30 open-pond cultivation cycles from a single field-study programme with one collection period per site, so the cross-site stability of Section 4.6 establishes robustness across environmental regimes but not across seasons, strains or cultivation systems. The missingness mechanism is characterized descriptively rather than tested. The simulated-missingness tier withholds values that were observed, so its error metrics describe reconstruction of observable cells and are informative about, but not identical to, the error incurred on the naturally missing cells, which remains unmeasurable by construction.

In this study, five possible extensions fall outside what the present dataset can support and are identified as future work. First is multi-seasonal and multi-strain replication requiring a different dataset, since the field study provides a single collection period per site. Second is formal missingness diagnostics, comprising a test of missingness completely at random and an explicit sensitivity analysis for missingness not at random, which would replace the descriptive characterization of Section 3.1. Third is formal significance testing of the differences between methods, which is deferred here, the coefficient of concordance reported in Section 4.4 establishing that the ranking is stable but not which pairs of methods are genuinely separated on each criterion. Fourth is deep-learning imputation baselines such as BRITS [23], SAITS [24] and GAIN [25], which fall outside the scope of this study and cannot in any case be trained or evaluated fairly at the sample size available, since the 30 production ponds provide 12,870 variable cells of which 8,171 are observed against 48,280 free parameters for BRITS at a 64-unit hidden state and 669,203 for SAITS at a model dimension of 256, giving 0.17 and 0.012 observed cells per free parameter against a rule-of-thumb requirement of roughly ten, while one sequence is one pond of 38 to 40 time-steps and only 30 exist, so that a 70/15/15 split leaves roughly 21 training, 4 validation and 5 test sequences. Fifth is a systematic multi-criteria comparison across decision-making families, using weights elicited from practitioners rather than the illustrative scenarios of Section 4.7, which would place the selection guideline on a firmer footing. Beyond these five, hybrid strategies that combine the mean-preserving and point-accuracy strengths of the ensemble and chained-equation learners with the dispersion-preserving properties of shape-preserving interpolation, of which the stochastic MICE variant evaluated here is a first step, remain the most promising route to simultaneous optimization across all validation dimensions.

5. Conclusion

This study evaluated twelve missing-data-handling methods for incomplete algal optical-density time-series across 30 open-pond cultivation cycles at five sites, using a two-tier validation framework that separates ground-truth-free assessment of the naturally missing observations from a supplementary simulated-missingness experiment.

To conclude, no single method is universally superior, the leading method changing with the criterion and autocorrelation preservation separating the methods least. MissForest and MICE jointly lead point reconstruction, at mean NRMSE of 0.1927 and 0.1933 under random masking and separated by less than one part in three hundred, with MICE taking the lead under contiguous outages (0.1841), but both compress variance and deviate distributionally, so their advantage is confined to point accuracy. PCHIP provides the best variance preservation, at a mean SD difference of 0.0056 and a dispersion ratio of 1.0009, and unlike the unconstrained cubic spline remains stable under multi-day outages, which makes it the appropriate interpolation choice where dispersion fidelity matters, while forward–backward fill matches the observed distribution most closely (0.2868) and imputes the largest share of missing values (90.11%) but is among the weakest on point accuracy.

The ranking is robust to re-randomization, at a coefficient of concordance of 0.955 across 30 masking configurations, and to change of site, but the design of the masking protocol matters, interpolation methods losing 2.4 to 3.0% of accuracy when gaps are contiguous rather than scattered while multivariate learners gain 3.4 to 4.8%. Ridge-regularized chained equations fitted over the whole cycle frame with a deterministic fallback eliminate the convergence failure of baseline MICE and raise coverage from 82.98% to 86.16%, adding stochastic residual draws reduces the distributional deviation from 0.5067 to 0.3454 at a measurable cost in point accuracy, and an aggregated analysis over four weighting scenarios selects polynomial fitting under equal, point-accuracy and temporal weightings and KNN under a variance weighting, which provides an explicit and auditable selection rule in place of a single global recommendation.

This study is limited with several limitation, comprising multi-seasonal and multi-strain replication beyond the single collection period available here, formal missingness diagnostics, significance testing of the method differences, deep-learning baselines on datasets large enough to train them, a broader multi-criteria comparison with elicited weights, and hybrid schemes that combine point-accuracy and dispersion-preserving behavior in a single imputer.

6. Declarations

6.1. Author Contributions

Conceptualization: E.J. and E.G.M.; Methodology: E.J., M.S.I.A.R. and A.A.M.; Software: E.J.; Validation: E.G.M., V.L. and A.F.; Formal Analysis: E.J. and L.G.P.W.; Investigation: E.J.; Writing - Original Draft: E.J.; Writing - Review and Editing: all authors; Supervision: E.G.M. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The ATP3 Unified Field Study dataset analyzed in this study is publicly available. The data presented in this study are available on request from the corresponding author. All additional tables can be found in the supplementary material document.

6.3. Funding

This work was supported by PETRONAS Research Sdn. Bhd. under the Algae Performance Prediction and Prescriptive Tools project, grant LKS2501. The authors also acknowledge the publication support provided by Penerbit UMS.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

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

References

- [1] P. L. Lien, T. T. Do, and T. Nguyen, "Data imputation for multivariate time-series data," in *Proc. Int. Conf. Knowledge Syst. Eng. (KSE)*, vol. 2023, no. Oct., pp. 1–6, Oct. 2023, doi: 10.1109/KSE59128.2023.10299484.
- [2] T. T. H. Phan, "Machine learning for univariate time series imputation," in *Proc. Int. Conf. Multimedia Anal. Pattern Recognit. (MAPR)*, vol. 2020, no. Oct., pp. 1–6, Oct. 2020, doi: 10.1109/MAPR49794.2020.9237768.
- [3] A. Chapon, T. B. Ouarda, and Y. Hamdi, "Imputation of missing values in environmental time series by D-vine copulas," *Weather Climate Extremes*, vol. 41, no. 1, pp. 1–10, 2023, doi: 10.1016/j.wace.2023.100591.
- [4] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, "A review of missing values handling methods on time-series data," in *Proc. Int. Conf. Inf. Technol. Syst. Innovation (ICITSI)*, vol. 2016, no. Oct., pp. 1–6, Oct. 2016, doi: 10.1109/ICITSI.2016.7858189.
- [5] S. Chitralekha, J. Prakash, H. Raghavan, R. Gopaluni, and S. Shah, "Comparison of expectation-maximization based parameter estimation using particle filter, unscented and extended Kalman filtering techniques," *IFAC Proc. Volumes*, vol. 42, no. 10, pp. 804–809, 2009, doi: 10.3182/20090706-3-FR-2004.00133.
- [6] O. Tamin, S. A. A. Karim, J. Sulaiman, E. G. Mounq, F. Khan, A. Baharum, and F. B. Muchtar, "Scattered data interpolation and approximation using machine learning: k-NN and DT approaches," *Lect. Notes Netw. Syst.*, vol. 2026, no. 1, pp. 193–204, 2026, doi: 10.1007/978-3-032-08246-6_16.
- [7] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018, doi: 10.1038/s41598-018-24271-9.
- [8] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012, doi: 10.1093/bioinformatics/btr597.
- [9] Y. Zhou, Q. Tang, and G. Zhao, "Gap infilling of daily streamflow data using a machine learning algorithm (MissForest) for impact assessment of human activities," *J. Hydrol.*, vol. 627, no. 1, pp. 1–12, 2023, doi: 10.1016/j.jhydrol.2023.130404.
- [10] H. Shin, T. Park, S.-K. Jo, and J. Y. Jung, "Enhancing flow-through aquaculture system monitoring: A comparative study of machine learning algorithms for missing-data imputation," *Aquaculture*, vol. 601, no. 1, pp. 1–12, May 2025, doi: 10.1016/j.aquaculture.2025.742303.
- [11] J. Park, K. Patel, and W. H. Lee, "Recent advances in algal bloom detection and prediction technology using machine learning," *Sci. Total Environ.*, vol. 938, no. 1, pp. 1–15, 2024, doi: 10.1016/j.scitotenv.2024.173546.
- [12] Y. Liu, B. Yang, K. Xie, J. Sun, and S. Zhu, "Dongting Lake algal bloom forecasting: Robustness and accuracy analysis of deep learning models," *J. Hazard. Mater.*, vol. 485, no. 1, pp. 1–12, 2025, doi: 10.1016/j.jhazmat.2024.136804.
- [13] J. Khoo, R. Alfred, K. Chekima, R. Pailus, C. K. On, E. G. Mounq, R. Alfred, O. V. Eboy, N. A. B. Raffie, A. O. Ibrahim, and N. Luan, "Enhancing carbon stock estimation using machine learning models: A comparative analysis of MLP, RNN, and autoencoder approaches," *Lect. Notes Netw. Syst.*, vol. 2025, no. 1, pp. 323–335, 2025, doi: 10.1007/978-981-96-6438-2_22.
- [14] E. Knoshaug, L. Laurens, C. Kinchin, and R. Davis, "Use of cultivation data from the Algae Testbed Public Private Partnership as utilized in NREL's algae state of technology assessments," *National Renewable Energy Laboratory*, Golden, CO, USA, Tech. Rep. NREL/TP-5100-67289, vol. 2016, no. 1, pp. 1–20, 2016, doi: 10.2172/1330992.
- [15] A. M. Hussein, M. Abd Elaziz, M. S. Abdel Wahed, and M. Sillanpää, "A new approach to predict the missing values of algae during water quality monitoring programs based on a hybrid moth search algorithm and the random vector functional link network," *J. Hydrol.*, vol. 575, no. 1, pp. 852–863, 2019, doi: 10.1016/j.jhydrol.2019.05.073.
- [16] W. Junger and A. Ponce de Leon, "Imputation of missing data in time series for air pollutants," *Atmos. Environ.*, vol. 102, no. 1, pp. 96–104, 2015, doi: 10.1016/j.atmosenv.2014.11.049.
- [17] M. Kazijevs and M. D. Samad, "Deep imputation of missing values in time series health data: A review with benchmarking," *J. Biomed. Inform.*, vol. 144, no. 1, pp. 1–12, 2023, doi: 10.1016/j.jbi.2023.104440.
- [18] N. Umar and A. Gray, "Comparing single and multiple imputation approaches for missing values in univariate and multivariate water level data," *Water*, vol. 15, no. 8, pp. 1–12, 2023, doi: 10.3390/w15081519.

-
- [19] I. Belachsen and D. M. Broday, "Imputation of missing PM2.5 observations in a network of air quality monitoring stations by a new kNN method," *Atmosphere*, vol. 13, no. 11, pp. 1–15, 2022, doi: 10.3390/atmos13111934.
- [20] O. Tamin, S. A. Abdul Karim, J. Sulaiman, E. G. Mounq, F. Khan, A. Baharum, and F. B. Muchtar, "Machine learning approaches in scattered data approximation: A comparative study of GPR and SVR," *Lect. Notes Netw. Syst.*, vol. 2026, no. 1, pp. 196–206, 2026, doi: 10.1007/978-3-032-08253-4_18.
- [21] D. Semou, D. Abdoulaye, H. D. El, F. Papa, and D. Ibrahima, "An evaluation of the performance of imputation methods for missing meteorological data in Burkina Faso and Senegal," *Afr. J. Environ. Sci. Technol.*, vol. 17, no. 10, pp. 252–274, 2023, doi: 10.5897/AJEST2023.3221.
- [22] L. E. Alejo-Sanchez, A. Márquez-Grajales, F. Salas-Martínez, A. Franco-Arcega, V. López-Morales, O. A. Acevedo-Sandoval, C. A. González-Ramírez, and R. Villegas-Vega, "Missing data imputation of climate time series: A review," *MethodsX*, vol. 15, no. 1, pp. 1–15, 2025, doi: 10.1016/j.mex.2025.103455.
- [23] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "BRITS: Bidirectional recurrent imputation for time series," *Adv. Neural Inf. Process. Syst.*, vol. 31, no. 1, pp. 6775–6785, 2018, doi: 10.48550/arXiv.1805.10572.
- [24] W. Du, D. Côté, and Y. Liu, "SAITS: Self-attention-based imputation for time series," *Expert Syst. Appl.*, vol. 219, no. 1, pp. 1–12, 2023, doi: 10.1016/j.eswa.2023.119619.
- [25] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," *Proc. Int. Conf. Mach. Learn.*, vol. 80, no. 1, pp. 5689–5698, 2018, doi: 10.48550/arXiv.1806.02920.