Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3953~3963
e-ISSN: 2548-6861

Performance of Multivariate Missing Data Imputation Methods on Climate Data

Amalia Safira Widyawati [1]*, Anwar Fitrianto [1]*, Pika Silvianti [1]*
* Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University
widyasafira@apps.ipb.ac.id [1]

Article Info

Article history:

Received 2025-09-23
Revised 2025-11-24
Accepted 2025-12-20

Keyword:

Average Per Month,
Climate Data,
Imputation,
K-NN,
Multivariate Lost Data

ABSTRACT

Climate data plays an important role in various aspects of life. However, missing data is often found, which can interfere with data processing and reduce the quality of analysis. Therefore, appropriate handling methods are needed to ensure that the analysis results remain valid. This study aims to compare the performance of several imputation methods for missing multivariate data based on the identification of actual missing data patterns, and to determine the appropriate imputation method based on the mechanism of missing data. This study also aims to apply the best method to data with actual missing data patterns to assess its effect on descriptive statistical changes required for further climatological analysis. The methods used include monthly averages, missRanger, k-Nearest Neighbor (k-NN), and Iterative Robust-Model Imputation (IRMI). The missing data information was obtained from Global Surface Summary of the Day (GSOD) data, namely temperature, precipitation, humidity, pressure, and wind speed variables with a daily frequency for 11 years, with a missing data proportion of 11.4%. The missing data patterns were then applied to relatively complete NASA Power data to evaluate the imputation results. The results show that IRMI is less capable of handling extreme missing data conditions, namely 17 completely missing rows. In contrast, k-NN, missRanger, and monthly averages provided better results in both extreme and non-extreme conditions. Of the four methods, monthly averages were chosen because they were able to overcome missing data while maintaining multivariate structure with 58% on sMAPE and 2.64% on relative difference.

I. PENDAHULUAN

Data iklim merupakan fondasi penting dalam perencanaan pertanian, mitigasi bencana, dan pengambilan kebijakan lingkungan di Indonesia. Informasi seperti curah hujan, suhu, kelembapan, dan kecepatan angin digunakan untuk memprediksi keadaan iklim harian maupun jangka panjang. Namun, analisis hanya akan menghasilkan kesimpulan yang valid apabila data yang digunakan berkualitas baik. Kualitas data yang baik dapat ditinjau dari beberapa dimensi, salah satunya adalah kelengkapan data (*completeness*). Ketidaklengkapan data menimbulkan kehilangan informasi (*information loss*), sehingga analisis hanya memanfaatkan sebagian informasi yang tersedia. Kondisi ini dapat menyebabkan hasil analisis menyimpang dari kondisi sebenarnya dan berisiko menghasilkan interpretasi yang tidak tepat [1].

Dalam praktiknya, data iklim yang tercatat sering tidak lengkap. Bahkan, pada beberapa stasiun pengamatan terdapat beberapa peubah seperti suhu yang sama sekali tidak tercatat selama beberapa tahun, sehingga informasi yang tersedia sangat terbatas untuk dianalisis. Hilangnya data iklim bisa disebabkan terjadinya kerusakan alat, gangguan teknis, kondisi cuaca ekstrem, ataupun keterbatasan sumber daya dalam proses pencatatan. Data hilang merupakan permasalahan umum yang terjadi dalam analisis data terutama pada kasus data iklim [2]. Selain mengurangi jumlah data, data hilang juga menimbulkan bias [3]. Hal ini menjadi masalah serius terutama dalam studi jangka panjang seperti

data iklim. Dengan kata lain, kualitas analisis data iklim sangat ditentukan oleh kelengkapan data yang tersedia.

Dampak tersebut terlihat nyata dalam berbagai aplikasi praktis, seperti penelitian di Bengkulu yang menunjukkan bahwa perhitungan debit puncak sungai dan simulasi banjir sepenuhnya bergantung pada data curah hujan [4]. Hal ini menegaskan pentingnya ketersediaan data curah hujan yang lengkap, karena jika tidak, estimasi debit maupun peta genangan yang dihasilkan menjadi bias dan tidak dapat diandalkan untuk pengambilan keputusan, sehingga mengurangi efektivitas perencanaan mitigasi banjir. Sementara itu, studi di Metropolitan Cali, Kolombia, menemukan bahwa meskipun data hilang hanya 0,5–5,4%, hal tersebut tetap berpotensi memengaruhi estimasi indeks kejadian curah hujan ekstrem [5]. Apabila data hilang tersebut diabaikan maka interpretasi tren iklim jangka panjang akan berisiko tidak tepat. Kedua contoh ini menunjukkan bahwa data hilang dalam proporsi kecil sekalipun dapat menimbulkan *information loss* dan berdampak serius dalam konteks yang berbeda.

Penanganan data hilang memerlukan pemahaman mendalam mengenai mekanisme data hilang serta metode yang tepat untuk mengatasinya. Little dan Rubin (2020) mengklasifikasikan mekanisme data hilang menjadi tiga, yaitu *Missing Completely At Random* (MCAR), *Missing At Random* (MAR), dan *Missing Not At Random* (MNAR). Menurut Gomer (2019), setiap mekanisme data hilang memerlukan metode penanganan yang berbeda pula. Oleh karena itu, penting untuk mengetahui mekanisme data hilang pada data sebelum melakukan penanganan. Selain mekanismenya**,** jenis metode imputasi juga perlu dipertimbangkan. Imputasi secara spesifik terbagi menjadi imputasi univariat dan imputasi multivariat. Menurut Templ dan Ulmer (2024), metode imputasi univariat seperti *mean imputation* cenderung memberikan hasil yang bias karena mengabaikan hubungan antar peubah. Kebalikan dari metode univariat, metode imputasi multivariat menggunakan hubungan antar peubah dalam memprediksi data yang hilang. Oleh sebab itu, pendekatan multivariat atau imputasi ganda umumnya lebih disarankan, terutama apabila data menghilang secara MAR.

Penelitian mengenai penanganan data hilang telah banyak dilakukan. Saeipourdizaj *et al.* (2021) membandingkan beberapa metode untuk mengimputasi data PM_{10} dan O_3. Penelitian tersebut menunjukan bahwa metode *k*-NN, interpolasi, dan *moving average* lebih baik dibandingkan *Predictive Mean Matching* (PMM). Selain itu, Templ dan Ulmer (2024) membandingkan sepuluh metode imputasi pada berbagai jenis data. Mereka menemukan bahwa metode *Iterative Robust-Model Imputation* (IRMI) unggul dalam menangani data hilang multivariat. Terakhir, Gurtskaia *et al.* (2024) melakukan simulasi penanganan data hilang menggunakan metode *Multiple Imputation by Chained Equations* (MICE), *Chained Random Forests with Predictive Mean Matching* (*missRanger* dengan PMM), dan *Extreme Gradient Boosting* pada data multi-level. Hasil penelitian menunjukkan bahwa metode terbaik untuk data multi-level adalah MICE dan *Extreme Gradient Boosting*.

k-Nearest Neighbor Imputation menjadi salah satu pendekatan yang banyak digunakan dalam menyelesaikan masalah imputasi karena keunggulannya saat dibandingkan metode lain. Algoritma ini menerapkan klasifikasi objek berdasarkan kumpulan data terdekat dengan menentukan jumlah tetangga terdekat (k). Penentuan jumlah k dilakukan berdasarkan perhitungan jarak antar tetangga dengan Euclidean *distance* [10]. Perhitungan jarak dilakukan dengan persamaan berikut:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

dengan p, q menunjukan titik data dan n jumlah atribut.

missRanger merupakan metode imputasi *machine learning* yang dikenalkan oleh Mayer (2024) dengan mengimplementasikan algoritma *Random Forest* dari paket *ranger.* Paket *ranger* merupakan optimasi dari algoritma *Random Forest* yang dipararelkan agar mempercepat waktu komputasi dan efisien dalam penggunaan memori [12]. Selain itu, dalam *missRanger* terdapat *Extremely Randomized Trees* (*ExtraTrees*) untuk meningkatkan keberagaman model. Metode ini memilih subset peubah dan ambang batas secara acak pada saat membangun pohon keputusan. Pendekatan ini akan menghasilkan pohon keputusan yang lebih bervariasi, sehingga meningkatkan kualitas imputasi *missing value* [13].

Metode *Iterative Robust-Model Imputation* merupakan metode imputasi yang menduga data hilang pada dataset secara iteratif menggunakan model regresi *robust* [14]. Menurut Templ (2023), metode ini efektif untuk menangani data multivariat yang hilang secara acak. Sesuai dengan namanya, algoritma IRMI tahan terhadap pencilan dan dapat menangani berbagai jenis peubah seperti kontinu maupun kategorik. Imputasi regresi klasik biasanya memerlukan setidaknya ada satu peubah tanpa *missing* agar bisa dijadikan prediktor yang stabil sedangkan IRMI lebih fleksibel karena tetap berjalan meskipun tidak ada satu pun peubah yang diamati sepenuhnya dalam dataset [16].

MCAR adalah kondisi ketika probabilitas suatu data hilang benar-benar terjadi secara acak dan tidak mengikuti pola tertentu [17]. Mekanisme MCAR tidak menimbulkan bias terhadap karakteristik data. Oleh karena itu, penanganan sederhana seperti *mean imputation* atau metode serupa masih dapat memberikan estimasi yang valid [18]. Mekanisme ini dapat diidentifikasi menggunakan uji Jamshidian dan Jalal (2010) melalui pendekatan pengujian homogenitas kovarian di antara kelompok-kelompok kasus dengan pola data hilang identik. Uji ini terdiri dari dua pendekatan utama yaitu pengujian normalitas dan homogenitas dengan uji Hawkins yang dimodifikasi dan pengujian homogenitas non-parametrik.

Berbeda dengan MCAR, mekanisme *Missing at Random* (MAR) justru ditandai dengan adanya pola. Pola ini berupa keterikatan antara peubah yang memiliki data hilang dengan peubah penjelas lainnya yang teramati [20]. Adanya hubungan antar peubah memerlukan perhatian dan banyak pertimbangan dalam proses analisis, karena pemilihan metode imputasi yang tidak sesuai akan menghasilkan estimasi yang bias [18]. Salah satu saran untuk memberikan penanganan pada mekanisme MAR yaitu imputasi ganda [18] dan pendekatan imputasi multivariat [8]. Imputasi ganda dinilai mampu memperhitungkan hubungan antar peubah dalam proses estimasi data hilangnya. Identifikasi mekanisme ini tidak dapat dilakukan melalui uji statistik formal seperti MCAR. Salah satu pendekatan yang umum digunakan untuk mengevaluasi kemungkinan MAR adalah melalui model regresi logistik [21].

MNAR adalah kondisi ketika probabilitas suatu data hilang bergantung pada data yang hilang itu sendiri. Hal tersebut membuat nilai yang hilang tidak diketahui secara pasti karena data yang hilang tidak dapat diamati sehingga sulit untuk ditangani secara statistik. Dalam praktiknya, membedakan antara MNAR dengan MAR cukup rumit karena kedua mekanisme tersebut bisa menunjukkan hubungan antara nilai yang hilang dengan nilai yang diketahui. Namun, yang membedakan MNAR adalah adanya hubungan dengan R dengan Y^m, yang tidak terdapat pada MAR [20]. Menurut Gomer (2020), MNAR tidak berbahaya jika dalam jumlah kecil. Namun masih sedikit metode yang terbukti menangani MNAR dengan memuaskan. Berbeda dengan MCAR dan MAR, MNAR tidak dapat diidentifikasi secara pasti. Hingga saat ini, belum ada metode statistik yang mampu membuktikan mekanisme MNAR secara langsung tanpa asumsi tambahan. Sehingga, pola hilang MNAR hingga saat ini hanya dapat dideteksi secara logis karena data yang hilang itu sendiri tidak tersedia dalam himpunan data [20].

Meskipun begitu, kebanyakan penelitian sebelumnya masih dilakukan dengan cara simulasi, yaitu menambahkan data hilang secara sengaja sesuai mekanisme MCAR, MAR, atau MNAR. Di sisi lain, studi yang menggunakan data iklim di Indonesia cenderung berfokus pada teknik imputasi data, tanpa membahas bagaimana data tersebut menghilang. Oleh karena itu, penting untuk terlebih dahulu mengenali pola data hilang yang sebenarnya sebelum menentukan metode imputasi yang tepat.

Penelitian ini bertujuan untuk membandingkan performa metode *mean* per bulan, *k*-NN, *missRanger*, dan IRMI dalam menangani data hilang multivariat berdasarkan pola data hilang aktual. Selain itu, penelitian ini juga bertujuan untuk menentukan metode imputasi yang sesuai berdasarkan mekanisme data hilang. Serta, penerapan metode terbaik pada data yang memiliki pola hilang aktual untuk menilai pengaruhnya terhadap perubahan statistik deskriptif yang diperlukan untuk analisis klimatologi lanjutan.

II. METODE

A. Data

Penelitian ini menggunakan dua data iklim, yaitu Global Surface Summary Of the Day (GSOD) dan NasaPower dengan kegunaan yang berbeda pada proses analisis. Data GSOD dipilih karena memiliki data hilang, sehingga dapat digunakan untuk mengetahui karakteristik data dan mengidentifikasi pola data hilang. Sementara itu, NasaPower yang relatif lengkap digunakan sebagai dasar simulasi imputasi. Pola data hilang yang diperoleh dari GSOD kemudian diterapkan pada NasaPower untuk mengevaluasi performa metode imputasi. Kedua data diambil dari titik yang sama, yaitu Stasiun Meteorologi Sultan Aji Muhammad Sulaiman Sepinggan, Kota Balikpapan, Provinsi Kalimantan Timur. Data tersebut terdiri dari delapan peubah iklim dengan amatan harian dari 1 Januari 2014 hingga 31 Desember 2024 atau selama 11 tahun, sehingga diperoleh sebanyak 4018 amatan berupa data harian. Deskripsi peubah ditunjukkan pada Tabel 1.

TABEL 1
PEUBAH YANG DIGUNAKAN

Kode	Keterangan	Satuan
Date	Tanggal	
TAvg	Suhu rata-rata	Celcius
TMax	Suhu maksimum	Celcius
TMin	Suhu minimum	Celcius
Prec	Curah hujan	Milimeter
Hr	Kelembapan	Persen
Press	Tekanan permukaan	MPa
Wind	Kecepatan angin	Kilometer/detik

B. Tahapan Analisis

Analisis data dilakukan menggunakan bantuan *software* Microsoft Excel dan Rstudio (versi 4.4.2). Paket yang digunakan pada proses analisis meliputi '*VIM'* untuk imputasi *k*-NN dan IRMI serta '*missRanger'* untuk imputasi berbasis *Random Forest.* Tahapan-tahapan analisis yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Mengunduh Data GSOD dan Data NasaPower.
2. Melakukan eksplorasi pada Data GSOD untuk melihat karakteristik data dan distribusi data hilang.
3. Mengidentifikasi pola data hilang pada Data GSOD berdasarkan hasil eksplorasi. Identifikasi ini dilakukan untuk menentukan apakah pola data hilang tersebut termasuk kategori MCAR, MAR, atau MNAR. Proses identifikasi setiap kategori mekanisme dilakukan dengan cara yang berbeda. I MCAR diidentifikasi dengan uji Jamshidian dan Jalal. MAR diidentifikasi menggunakan regresi logistik untuk setiap pebah yang memiliki nilai hilang dan peubah lain sebagai peubah prediktor. Peubah yang hilang diubah menjadi peubah *dummy*

dengan 1 jika observasi tersedia dan 0 jika data hilang. Peubah *dummy* tersebut menjadi peubah respon pada Dan identifikasi MNAR ditentukan berdasarkan eleminasi dari identifikasi MCAR dan MNAR serta didukung studi literatur.

4. Menerapkan pola data hilang dari GSOD pada NasaPower untuk menguji kebaikan metode imputasi.
5. Melakukan pendugaan data hilang menggunakan metode *mean* per bulan pada Data NasaPower.
6. Melakukan pendugaan data hilang menggunakan metode *k*-NN menggunakan pada Data NasaPower. Langkah-langkah yang dilakukan dalam *k*-NN meliputi:
   a. Menentukan parameter *k*. Penentuan *k* dilakukan secara empiris dari 1-25.
   b. Menghitung jarak Euclidean
   c. Mengurutakan hasil perhitungan jarak Euclidean dan memiliki *k* tetangga terdekat.
   d. Menghitung nilai *weight mean estimation* sesuai jumlah *k* yang telah ditentukan sebelumnya. Dengan persamaan *weight mean estimation* sebagai berikut:

$$x_j = \frac{\sum_{K=1}^{K} w_k v_k}{\sum_{K=1}^{K} w_k} \quad (10)$$

dengan nilai pengamatan tetangga terdekat ke-*k*

$$w_k = \frac{1}{d_{(x,y)}{}^2} \quad (11)$$

Keterangan:
k : jumlah parameter *k* yang digunakan
v_k : bobot berdasarkan data lengkap pada atribut yang mengandung data hilang berdasarkan parameter *k*
$d_{(x,y)}$: jarak *Euclidean* antar parameter *k*

7. Melakukan pendugaan data hilang menggunakan metode IRMI pada Data NasaPower. Langkah-langkah yang dilakukan dalam IRMI meliputi:
   a. Menginisiasi awal dengan mengisi data hilang menggunakan metode imputasi sederhana (seperti median sebagai *default*).
   b. Pengurutan peubah bedasarkan jumlah data hilang aslinya, dari yang paling banyak hingga paling sedikit.
   c. Peubah yang telah diurutkan akan dijadikan peubah respon dan peubah lain menjadi prediktor.
   d. Koefisien regresi diestimasi dan data yang hilang pada peubah respon tersebut diimputasi menggunakan model yang telah dilatih.
   e. Proses c dan d diulang hingga data yang diimputasi stabil atau mencapai jumlah iterasi maksimum.
8. Melakukan pendugaan data hilang menggunakan metode *missranger* pada Data NasaPower. Langkah-langkah yang dilakukan dalam missRannger meliputi:
   a. Membagi data menjadi data lengkap dan data hilang.
   b. Membangun model *Random Forest* dengan peubah lain sebagai predictor
   c. Menggunakan model yang telah dilatih untuk memprediksi nilai yang hilang
   d. Membuat kumpulan nilai imputasi berdasarkan prediksi dari model
9. Mengevaluasi dan menentukan hasil prediksi pendugaan data hilang dengan menghitung *Mean Absolute Error* (MAE), *Symmetric Mean Absolute Percent Error* (sMAPE), *Root Mean Squared Error* (RMSE), *relative difference* dan visualisasi menggunakan *scatter plot with ellips* 95% untuk mendapatkan metode terbaik. Masing-masing metrik evaluasi dinotasikan sebagai berikut:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t{}^2}$$

$$sMAPE = \frac{1}{n}\sum_{t=1}^{n} \frac{2 \cdot |e_t|}{|Y_t| + |F_t|}$$

$$rell.diff.\left(\boldsymbol{B}_{[\boldsymbol{p}\times\mathbf{2}]}, \boldsymbol{B}_{[\boldsymbol{p}\times\mathbf{2}]}^{(imp)}\right) = \frac{1}{2p}\sum_{j=1}^{2}\sum_{i=1}^{p}\left|\frac{b_{ij} - b_{ij}^{(imp)}}{b_{ij}}\right|$$

Keterangan:
F_t : nilai prediksi hasil imputasi pada NasaPower (setelah diterapkan pola hilang)
Y_t : nilai aktual pada data lengkap NasaPower (sebelum diterapkan pola hilang)
n : jumlah total data
e_t : selisih antara nilai aktual (Y_t) dan nilai prediksi (F_t)
$\boldsymbol{B}_{[\boldsymbol{p}\times\mathbf{2}]}$: matriks loading *Principal Component Analysis* (PCA) nilai aktual

$\boldsymbol{B}^{(imp)}_{[p\times 2]}$: matriks loading PCA nilai prediksi
i : indeks peubah
b_{ij} : nilai loading peubah ke-i pada komponen utama ke-j untuk data aktual.
$b^{(imp)}_{ij}$: nilai loading peubah ke-i pada komponen utama ke-j untuk data prediksi.

10. Menerapkan metode terbaik pada Data GSOD serta membandingkan statistik deskriptif sebelum dan sesudah imputasi.

III. Hasil dan Pembahasan

A. *Eksplorasi Data GSOD*

1. Statistik deskriptif

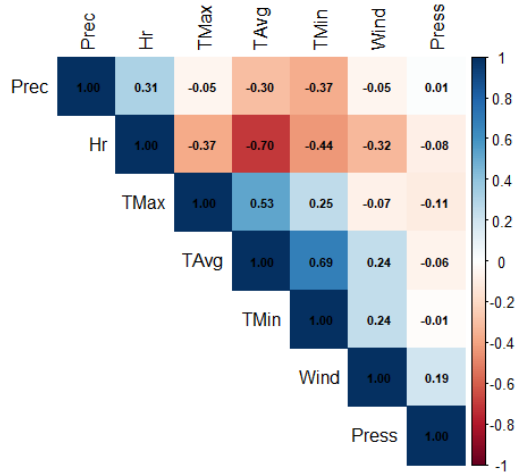
Statistik deskriptif membantu mengidentifikasi data menjadi informasi yang lebih detail dan mudah dipahami. Berdasarkan Tabel 2, terlihat setiap peubah memiliki data yang hilang yang bervariasi.

TABEL 2
STATISTIK DESKRIPTIF DATA GSOD

Kode	Jumlah data hilang	*Mean*	Median	Deviasi Standar
TAvg	30	27,8	27,9	0,9
TMax	24	31,2	31,4	1,1
TMin	30	24,7	24,5	0,9
Prec	450	10,8	1,1	23,5
Hr	19	82,7	82,9	4,9
Press	19	6,8	6,0	3,1
Wind	19	1009,4	1009,4	1,3

2. Korelasi antar Peubah

Korelasi dapat memberikan informasi berupa hubungan antar peubah untuk memperkirakan adanya suatu hubungan sebab-akibat apabila terdapat teori, logika, maupun observasi yang dapat dipertanggungjawabkan.



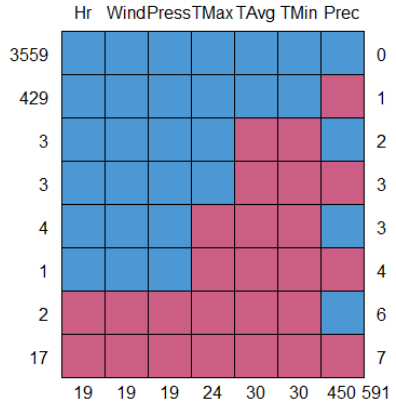
Gambar 1.Korelasi antar peubah

Gambar 1 menunjukkan bahwa TAvg memiliki pengaruh yang besar terhadap peubah TMax, TMin, dan Hr. Hal ini sesuai dengan prinsip Clausius–Clapeyron [22] yang menyatakan bahwa suhu tinggi meningkatkan kapasitas maksimum udara untuk menahan uap air. Namun, apabila kandungan uap air aktual tidak bertambah maka kelembapan relatif cenderung menurun.

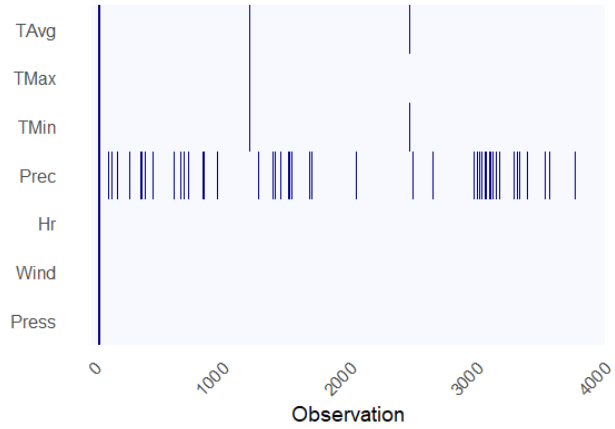
Sebaliknya, peubah Prec, Wind, dan Press memperlihatkan nilai korelasi yang rendah terhadap peubah lain. Hubungan yang lemah ini kemungkinan besar dipengaruhi oleh faktor eksternal yang kompleks dan tidak secara langsung. Misalnya, penelitian Qi et al. (2023) menunjukkan bahwa curah hujan berkorelasi rendah dengan kecepatan angin, di mana kecepatan angin meningkat dua hari sebelum hujan dan menurun setelah hujan terjadi. Hal ini menunjukkan adanya hubungan sebab-akibat yang memiliki jeda waktu. Selain itu, faktor eksternal seperti region dan *El Niño–Southern Oscillation* (ENSO) diketahui berperan penting dalam membentuk variabilitas iklim regional. Xu *et al.* (2023) menunjukkan bahwa hubungan antara sistem monsun Asia-Australia dan ENSO semakin menguat sejak pertengahan abad ke-19, yang diduga dipengaruhi oleh pemanasan global. Pengaruh eksternal ini dapat berkontribusi pada kompleksitas hubungan antar peubah iklim lokal yang diamati.

3. Visualisasi Data Hilang

Keberadaan data hilang pada data iklim merupakan isu penting dalam pengolahan data. Pada Tabel 2 ditampilkan bahwa data iklim Balikpapan memiliki sejumlah data hilang pada beberapa peubah. Untuk mengidentifikasi pola dan distribusi hilang digunakan fungsi *md.pattern* dari paket *mice* yang menunjukkan setiap peubah serta jumlah nilai yang hilang pada masing-masing observasi, seperti yang ditampilkan pada Gambar 2.



Gambar 2.Pola data hilang berdasarkan peubah



Gambar 3.Pola data hilang berdasarkan hari

Angka pada sumbu *y* menunjukan jumlah data yang hilang pada kolom berwarna merah dan angka pada sumbu *x* menunjukan jumlah data yang hilang pada masing-masing peubah. Berdasarkan Gambar 2, diketahui bahwa data hilang terbanyak terdapat pada peubah Prec, yaitu sebanyak 450 amatan. Rincian pola data hilang dalam periode 11 tahun ditampilkan pada Gambar 3. Pola data hilang peubah Prec lebih bervariasi dibandingkan dengan peubah lainnya yang cenderung konstan.

Tingginya jumlah data hilang pada peubah Prec dapat disebabkan oleh variabilitas spasiotemporal pada curah hujan yang tinggi. Penelitian oleh Aguilera *et al.* (2020) menyatakan bahwa curah hujan merupakan peubah yang sulit ditangani karena memiliki keragaman yang tinggi dalam ruang dan waktu. Hal tersebut menjadikan peubah Prec sering menghasilkan pola data yang hilang secara acak dan kronologis dalam rangkaian waktu. Masalah data hilang dalam data hidrometeorologi lainnya disebabkan oleh banyak hal, di antaranya kesalahan pengukuran, kerusakan pada sensor, kesalahan dalam akuisisi dari operator, ketidakaktifan sementara stasiun, dan perubahan konfigurasi jaringan.

B. *Identifikasi Pola Data Hilang pada Data GSOD*

1) *MCAR*

Pengujian mekanisme data hilang dilakukan menggunakan paket *missMech* pada R, sesuai metode yang dikembangkan oleh Jamshidian dan Jalal (2010).

TABEL 3
HASIL UJI MCAR

Jenis Uji	*p-value*
Hawkin's Test	0,00*
Non-parametric Test	0,02*

Berdasarkan Tabel 3, kedua hasil menunjukkan nilai *p-value* <0,05 yang berarti menolak hipotesis nol bahwa data hilang terjadi secara acak sempurna. Dengan demikian, data hilang GSOD tidak termasuk dalam kategori MCAR. Penolakan hipotesis MCAR ini konsisten dengan temuan sebelumnya bahwa curah hujan memiliki karakteristik kompleks dengan variabilitas tinggi terhadap ruang dan waktu. Kondisi ini mengindikasikan adanya faktor eksternal yang tidak acak, seperti kondisi ekstrem atau kegagalan pencatatan saat hujan tinggi. Oleh karena itu, diperlukan pengujian lebih lanjut untuk mengevaluasi kemungkinan bahwa data menghilang secara MAR atau MNAR.

2) *MAR*

Pengujian MAR dilakukan melalui pendekatan eksploratif dengan menggunakan model regresi logistik untuk setiap peubah respon yang memiliki data hilang.

TABEL 4
HASIL REGRESI LOGISTIK UNTUK SETIAP PEUBAH

Peubah Respon *Dummy*	Signifikansi
TAvg	Tidak ada
TMax	Tidak ada
TMin	Tidak ada
Prec	Ada
Hr	Tidak ada
Wind	Tidak ada
Press	Tidak ada

Berdasarkan Tabel 4, hanya peubah curah hujan yang menunjukkan hubungan signifikan dengan peubah lainnya. Hasil ini mengindikasikan bahwa data menghilang secara MAR. Hal ini bisa dikaitkan dengan sifat curah hujan yang memiliki keragaman antara ruang dan waktu sehingga data hilang rawan terjadi. Selain itu, bisa disebabkan oleh gangguan pengukuran, kesalahan akuisisi, atau kerusakan sensor yang cenderung terjadi pada kondisi cuaca ekstrem. Dengan demikian, data hilang pada peubah curah hujan dapat dikaitkan dengan peubah lain yang diamati, sehingga mendukung asumsi MAR. Sementara itu, percobaan regresi logistik pada peubah lainnya seperti TAvg, Press, Wind, Hr, TMin, dan TMax tidak menunjukan hubungan signifikan terhadap peubah lainnya. Oleh karena itu, tidak terdapat cukup bukti untuk menyatakan bahwa peubah-peubah tersebut mengikuti mekanisme MAR.

3) *MNAR*

Berdasarkan hasil evaluasi asumsi MCAR dan MAR, kemungkinan besar data hilang pada peubah TAvg, Press, Wind, Hr, TMin, dan TMax tidak terjadi secara acak sempurna dan tidak dapat dijelaskan oleh peubah lain yang teramati. Dengan demikian, peubah-peubah tersebut diasumsikan mengikuti mekanisme MNAR melalui pendekatan eliminasi terhadap MCAR dan MAR. Zhou *et al.* (2024) membagi MNAR menjadi dua kategori yang lebih rinci yaitu *Focused* MNAR dan *Diffuse* MNAR. *Focused* MNAR merupakan situasi data hilangnya hanya bergantung pada nilai yang hilang (Y^m) dan tidak bergantung pada data yang teramati (Y^o).

Sementara itu, *Diffuse* MNAR merupakan situasi ketika proses hilangnya data tersebut melibatkan nilai yang hilang (Y^m) maupun nilai yang teramati (Y^O).

Berdasarkan pola data hilang yang bersifat kronologis selama 17 hari berturut-turut, serta karakteristik statistik peubah yang dianalisis, mekanisme *Diffuse MNAR* dinilai lebih relevan dalam konteks penelitian ini. Pola data hilang tersebut tidak mungkin bersifat acak sepenuhnya, dan juga tidak dapat dijelaskan oleh kejadian iklim ekstrem jangka pendek. Hal ini menunjukkan kemungkinan besar terdapat interaksi antara karakteristik intrinsik data dan faktor eksternal. Namun demikian, karena mekanisme MNAR tidak dapat diidentifikasi secara pasti melalui metode statistik saja, diperlukan pendekatan kontekstual, pertimbangan ahli, dan analisis sensitivitas sebagaimana disarankan oleh Mason *et al.* (2020).

C. Hasil *Pendugaan* Data Hilang pada Data NasaPower

Setelah pola data hilang diidentifikasi pada GSOD, tahap berikutnya adalah melakukan pendugaan data hilang melalui proses imputasi. Pendugaan data hilang merupakan proses penting ketika terdapat data yang hilang pada peubah iklim. Dalam studi ini, dilakukan imputasi terhadap setiap peubah yang memiliki data hilang. Proses imputasi dilakukan pada Data NasaPower, yang sebelumnya telah disesuaikan dengan pola data hilang pada GSOD. Tujuannya adalah agar hasil imputasi dapat dievaluasi dengan membandingkan hasil prediksi terhadap nilai data lengkap yang diketahui.

Metode imputasi yang digunakan meliputi mean per bulan, k-Nearest Neighbor (k-NN), missRanger, dan IRMI. Subbab ini menyajikan beberapa hasil imputasi dari masing-masing metode pada data dengan baris yang sepenuhnya hilang untuk peubah Prec (curah hujan). Peubah ini dipilih karena memiliki jumlah data hilang yang lebih banyak dibandingkan dengan peubah lainnya.

TABEL 5
NILAI METRIK EVALUASI HASIL IMPUTASI

Tanggal	Nilai Data	*Mean* per bulan	*k*-NN	*miss-Ranger*
29/10/2024	1,65	4,3	6,69	4,5
30/10/2024	1,26	4,3	6,69	4,0
31/10/2024	1,63	4,3	6,69	4,5
01/11/2024	2,68	8,8	6,69	4,6
02/11/2024	2,88	8,8	6,69	4,7

Berdasarkan Tabel 5, setiap metode memberikan hasil pendugaan yang sangat beragam. *Mean* per bulan pada bulan Oktober menunjukan selisih yang cukup besar dibandingkan dengan hasil pendugaan metode lain. Hasil pendugaan *k*-NN pada *k* terbaik *k*=23 memberikan hasil yang identik selama lima hari berturut-turut. Hal ini menunjukan bahwa kedua metode ini tidak mencerminkan variasi harian curah hujan. Hasil pendugaan metode *missRanger* memberikan hasil yang berbeda setiap harinya. Meskipun selisih antara hasil pendugaan dan nilai lengkapnya masih cukup besar, namun metode ini mampu merepresentasikan fluktuasi curah hujan harian dengan lebih baik. Sedangkan IRMI tidak dapat memberikan *output* karena adanya 17 hari hilang seluruhnya berturut-turut. Kondisi ini menjadi kondisi ekstrem karena hilangnya seluruh baris data dalam periode waktu yang panjang dan berkelanjutan.

D. Evaluasi *Hasil* Pendugaan

1) Evaluasi Numerik

Akurasi dan evaluasi model dilakukan sebagai tolak ukur untuk menentukan metode terbaik dalam imputasi data iklim. Hasil imputasi yang diperoleh akan diuji menggunakan metrik evaluasi, yaitu MAE, RMSE, dan sMAPE. Ketiga metrik ini digunakan untuk melihat seberapa besar perbedaan antara nilai hasil imputasi dengan data lengkap. Semakin kecil nilai dari ketiga metrik tersebut, maka semakin baik metode tersebut dalam mengisi data yang hilang.

TABEL 6
NILAI METRIK EVALUASI HASIL IMPUTASI

Metode	MAE	RMSE	sMAPE (%)
Mean per bulan	3,2	4,9	58,0
k-NN	2,7	4,7	53,0
missRanger	2,9	4,7	55,6

Berdasarkan Tabel 6, terlihat bahwa metode *k*-NN menghasilkan nilai MAE, RMSE, dan sMAPE paling rendah. Hal ini menunjukkan bahwa metode ini memiliki performa yang sama baiknya pada kondisi ekstrem. *missRanger* dan *mean* per bulan masih cukup kompetitif dan mendekati kinerja *k*-NN. Adapun IRMI tidak menghasilkan *output* karena metode ini gagal dijalankan akibat adanya baris data yang sepenuhnya hilang.

Berdasarkan penelitian yang dilakukan oleh Templ *et al.* (2011), IRMI seharusnya mampu menangani data dengan data hilang multivariat, bahkan saat tidak ada peubah yang sepenuhnya teramati. Namun, pada penelitian ini, IRMI gagal dijalankan karena keberadaan 17 baris yang hilang keseluruhan. Hal ini menunjukkan adanya keterbatasan dalam fungsi *irmi*() untuk mencapai proses inisiasi atau konvergensi akibat kompleksitas pola hilang pada studi ini. Untuk memastikan hal tersebut dilakukan percobaan dengan mengimputasi pada salah satu peubahnya dan memberikan hasil IRMI dapat berjalan dengan baik. Temuan ini menunjukan bahwa IRMI membutuhkan setidaknya satu peubah untuk mampu menginisiasi atau mencapai konvergensi. Selain itu, sejalan dengan penelitian Resheff dan Weinshall (2017) yang menunjukkan bahwa IRMI dapat gagal mencapai konvergensi bahkan pada dataset dengan hanya 5% data yang hilang secara acak, seperti pada kasus data *wine* dan *storks*. Hal ini memperkuat fungsi *irmi()* memiliki keterbatasan pada data dengan kondisi tertentu.

Menurut Templ (2023), IRMI lebih optimal diterapkan pada data multivariat dengan mekanisme MAR. Berdasarkan algoitma, IRMI merupakan metode berbasis regresi yang membutuhkan peubah lain untuk mendapatkan nilai imputasi. Secara teoritis, algoritma ini tidak sesuai dengan mekanisme MNAR yang hilang karena nilai hilang itu sendiri. Hal ini konsisten dengan temuan penelitian ini, di mana IRMI menunjukkan ketidakmampuan dalam menangani baris sepenuhnya hilang yang berpola MNAR.

Skenario kedua dibuat dengan menghapus 17 baris yang hilang sepenuhnya agar metode IRMI dapat dijalankan. Setelah penghapusan, pola hilang yang tersisa didominasi pola MAR pada peubah curah hujan. Namun, sebagian peubah lain masih mungkin mengikuti pola MNAR karena hilangnya tidak sepenuhnya dapat dijelaskan oleh peubah lain. Dengan demikian, skenario kedua dapat dianggap sebagai uji sensitivitas untuk menilai kinerja metode imputasi ketika setiap baris data menyisakan minimal satu nilai teramati.

TABEL 7
NILAI METRIK EVALUASI HASIL IMPUTASI TANPA POLA HILANG KESELURUHAN

Metode	MAE	RMSE	sMAPE (%)
Mean per bulan	3,7	5,4	67,9
k-NN	2,9	5,1	62,0
missRanger	3,4	5,1	65,5
IRMI	14,6	30,9	94,9

Pada Tabel 7, terlihat bahwa metode *k*-NN kini menunjukkan kinerja terbaik dengan nilai MAE, RMSE, dan sMAPE yang paling rendah. Metode *missRanger* dan *mean* per bulan juga tetap menunjukkan performa yang cukup baik dan stabil pada kedua skenario. Meskipun IRMI berhasil dijalankan, metode ini menghasilkan kesalahan prediksi terbesar pada seluruh metrik. Hal ini menunjukkan bahwa keunggulan teoretis IRMI untuk data hilang multivariat dan ketahanannya terhadap pencilan kurang optimal ketika diterapkan pada data dalam studi ini.

Karakteristik data iklim yang digunakan dalam penelitian ini memiliki keterkaitan erat dengan performa masing-masing metode imputasi. Templ *et al.* (2011) menyatakan bahwa kesalahan prediksi imputasi akan cenderung lebih kecil apabila hubungan antar peubah tinggi. Sebaliknya, peubah dengan korelasi lemah menghasilkan prediksi yang buruk dan lebih banyak *noise*. Rendahnya nilai korelasi dapat memengaruhi menurunnya akurasi model imputasi berbasis regresi linier seperti IRMI.

Selanjutnya, metode *k*-NN kembali menunjukan performa yang sangat baik pada skenario kedua ini. Hasil ini memperlihatkan bahwa *k*-NN mampu mengatasi keadaan yang ekstrem maupun tidak. Kestabilan ini diperkuat oleh penelitian Murti *et al.* (2019) yang menyatakan bahwa *k*-NN mampu mengimputasi dengan cukup baik bahkan dengan kasus data hilang sebesar 20%. Hal ini menunjukkan bahwa *k*-NN tahan terhadap pola data hilang yang berbeda dan tetap menghasilkan imputasi yang mendekati data lengkap pada kondisi ekstrem. Berbeda dengan metode *eager learner* yang membentuk model, *k*-NN merupakan *lazy learner* yang memprediksi berdasarkan kedekatan jarak antar amatan dan tidak bergantung pada asumsi model (Han dan Kamber 2001). Karakteristik ini membuat *k*-NN mampu memberikan hasil evaluasi yang baik meskipun terdapat pencilan ekstrem dan data bersifat nonlinier. Selain itu, berdasarkan hasil evaluasi, *k*-NN terbukti efektif dengan data yang mengandung pola hilang MNAR dan MAR. Penemuan ini diperkuat oleh penelitian [28] yang menyatakan bahwa *k*-NN konsisten memberikan hasil yang baik pada berbagai mekanisme.

Di sisi lain, *missRanger* menunjukkan performa yang cukup stabil pada kedua skenario. Namun, penurunan performa pada skenario kedua mengindikasikan bahwa model ini membutuhkan keragaman data untuk membangun pola adaptif. Homogenitas data akibat penghapusan baris membatasi struktur yang diperlukan pohon keputusan, sehingga menurunkan akurasi imputasi. Berdasarkan hasil kedua skenario pola hilang yang berbeda tersebut, *missRanger* terbukti fleksibel baik pada kondisi MAR maupun MNAR.

Metode *mean* per bulan menunjukkan performa yang relatif stabil pada kedua skenario, meskipun masih berada di bawah *missRanger*. Namun, imputasi ini lebih baik dibandingkan rata-rata keseluruhan karena mampu mempertahankan pola musiman dalam data iklim. Secara umum, hasil amatan data GSOD dan Nasapower memiliki perbedaan. Data Nasapower diambil menggunakan titik koordinat yang sama dengan lokasi Stasiun BMKG Balikpapan seperti GSOD. Meskipun begitu, Nasapower memiliki data yang lebih *smooth* karena diambil menggunakan satelit dan resolusi grid sedangkan GSOD menggunakan amatan titik secara langsung. Oleh karena itu pada penelitian ini *mean* per bulan memiliki kemampuan yang hampir sama baiknya dengan *machine learning*. Metode ini lebih sesuai diterapkan pada data dengan mekanisme MCAR karena sifatnya yang sederhana. Oleh karena itu, perlu dilakukan uji coba kembali apabila metode ini akan diterapkan pada dataset lain dengan karakteristik berbeda.

Selain itu, secara geografis, Balikpapan berada dekat dengan garis khatulistiwa dan berada di pesisir timur Kalimantan. Berdasarkan klasifikasi Köppen-Geiger, wilayah khatulistiwa seperti Balikpapan diklasifikasikan sebagai Iklim Tropis Hutan Hujan (Af) [29]. Fenomena ini didukung oleh data BMKG, yang menyatakan bahwa zona iklim khatulistiwa memiliki curah hujan yang tinggi dan suhu yang konstan sepanjang tahun [30]. Kestabilan iklim yang dominan siklus tahunan ini menghasilkan kestabilan antar bulan yang dapat ditangkap secara efektif hanya

dengan menggunakan *mean* per bulan. Hal ini diperkuat dengan data statistik deskriptif pada Tabel 2, setiap peubah memiliki nilai deviasi standar yang kecil kecuali pada curah hujan karena variasi hujan yang memang cukup tinggi pada daerah ini.

TABEL 8
HASIL UJI POST-HOC DUNN

Comparison	P.adj
k-NN_*Mean* per bulan	$7{,}2e^{12}$
k-NN_*missRanger*	$8{,}2e^{03}$
*missRanger*_*mean* per bulan	$4{,}4e^{23}$

Hasil uji Friedman memberikan *p-value* sebesar $2{,}2e^{16}$ dengan $\alpha = 0{,}05$ membuktikan terdapat perbedaan kinerja akurasi akurasi numerik antar metode. Lebih lanjut, uji *Post-Hoc* Dunn mengonfirmasi bahwa semua pasangan metode menunjukan perbedaan kinerja signifikan secara statistik yang ditunjukan pada Tabel 8. Berdasarkan nilai median RMSE yang paling rendah, metode *k*-NN ditetapkan sebagai metode yang paling unggul secara numerik, diikuti oleh *missRanger* dan *mean* per bulan.

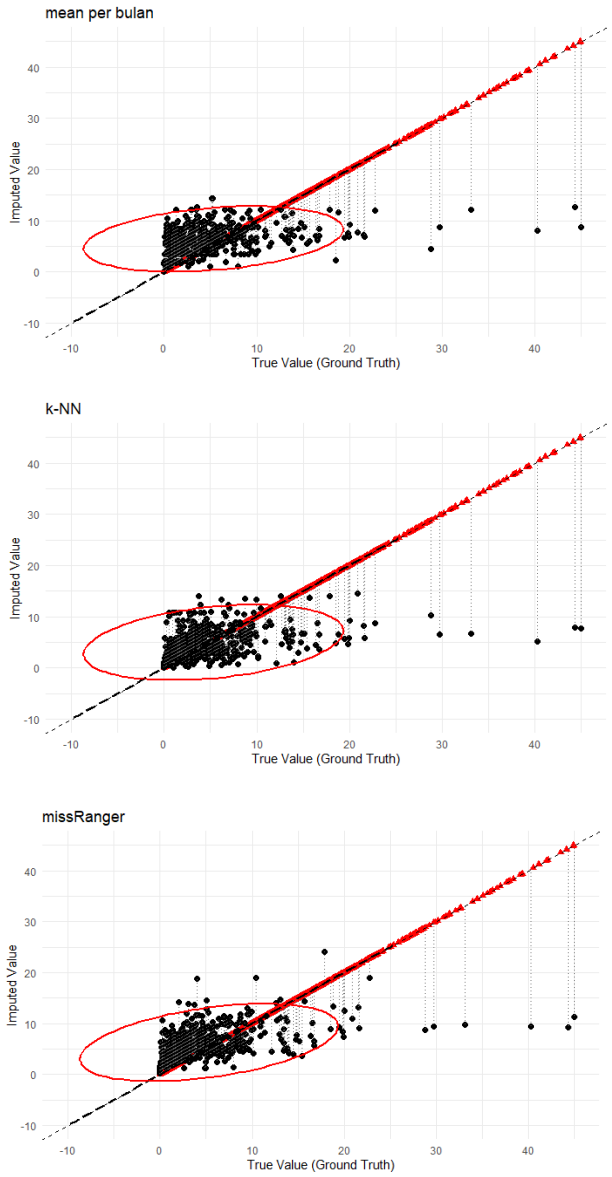
2) Evaluasi Visual

Berdasarkan hasil uji *post-hoc* Dunn, terdapat perbedaan kinerja RMSE antar pasangan metode imputasi. Secara numerik, metode *k*-NN dan *missRanger* menunjukan performa yang lebih unggul dibandingkan dengan *mean* per bulan dalam mengestimasi nilai imputasi. Namun, meskipun secara numerik lebih rendah, nilai metrik evaluasi masih dalam rentang yang kompetitif. Sehingga ketiga metode tersebut layak untuk dipertimbangkan lebih lanjut.

Selain evaluasi numerik, pemilihan metode kemudian dilakukan dengan evaluasi visual. Evaluasi visual dilakukan pada skenario pertama menggunakan *scatter plot with ellips* 95% dan nilai *relative difference* untuk menilai kemampuan masing-masing metode dalam mempertahankan struktur multivariat data. Pemilihan metode pada tahap ini lebih dicondongkan pada nilai *relative differencei* terkecil, yang mengindikasikan stuktur hasil imputasi semakin dekat dengan struktur data lengkap. Selain itu, *scatter plot* juga digunakan untuk memverifikasi apakah titik-titik hasil imputasi mendekati garis diagonal, yang menunjukkan kedekatan nilai imputasi dengan nilai aslinya.

TABEL 9
NILAI *RELATIVE DIFFERENCE* SETIAP METODE

Metode	*Relative difference* (%)
Mean per bulan	2,64
k-NN	7,57
missRanger	7,33



Gambar 4. Evaluasi visual setiap metode

Berdasarkan Gambar 4, terlihat bahwa titik-titik hitam pada metode *mean* per bulan cenderung mengumpul di dalam elips merah. Hal ini menunjukkan bahwa hasil imputasi dengan metode ini cukup mendekati nilai aslinya sekaligus menjaga struktur data secara umum. Selain itu, nilai *relative difference* yang rendah pada Tabel 9 memperkuat bahwa metode ini mampu mempertahankan struktur multivariat data, khususnya arah dan besaran vektor pemuatan pada dua komponen utama PCA.

Sementara itu, sebaran titik pada metode *k*-NN cenderung lebih terpusat di dalam elips merah dibandingkan *missRanger*. Hal ini menunjukkan bahwa distribusi amatan dari hasil imputasi dengan *k*-NN lebih terkonsentrasi. Namun, nilai *relative difference* pada *missRanger* lebih rendah dibandingkan pada *k*-NN. Meskipun secara visual sebaran

hasil imputasi *missRanger* terlihat lebih luas, nilai *relative difference* yang lebih rendah menunjukkan bahwa metode ini lebih mampu mempertahankan struktur multivariat data.

Temuan ini menunjukkan adanya ketidaksesuaian antara hasil evaluasi numerik dan visual. Secara numerik, metode *k*-NN memberikan nilai error paling rendah. Namun, dari sisi pelestarian struktur data multivariat, *mean* per bulan justru lebih unggul. Hal ini dapat dijelaskan karena metode *mean* per bulan mengimputasi data berdasarkan rata-rata historis musiman, sehingga tidak mengubah pola distribusi dan relasi antar peubah secara drastis. Sementara itu, metode seperti *k*-NN yang menggunakan kedekatan jarak bergantung pada keragaman data latih yang tersedia. Sehingga, memengaruhi sebaran data multidimensi setelah imputasi.

Dengan demikian, evaluasi visual ini memberikan gambaran tambahan yang tidak sepenuhnya tercermin dalam evaluasi numerik. Kombinasi kedua pendekatan ini penting untuk memberikan pemahaman yang menyeluruh terhadap efektivitas metode imputasi yang diuji. Dalam konteks penelitian ini, metode *mean per bulan* terbukti mampu mempertahankan struktur multivariat data yang tercermin dari nilai *relative difference* yang rendah, sementara *k*-NN unggul dalam fleksibilitas penyesuaian terhadap kondisi data hilang yang lebih kompleks. Oleh karena itu, pemilihan metode imputasi yang tepat perlu mempertimbangkan tidak hanya nilai kesalahan prediksi. Aspek lain yang juga penting adalah kemampuan mempertahankan struktur internal data, terutama saat digunakan untuk analisis multivariat seperti PCA atau analisis klasterisasi.

E. Penerapan pada Data *GSOD*

Pada tahap ini dilakukan penerapan metode imputasi terhadap Data GSOD dengan menggunakan metode *mean* per bulan sebagai metode optimal berdasarkan hasil evaluasi pada tahap sebelumnya. Untuk menilai sejauh mana metode imputasi ini memengaruhi distribusi data, dilakukan perbandingan statistik deskriptif antara data sebelum dan sesudah imputasi. Statistik deskriptif yang digunakan meliputi nilai *mean*, median, standar deviasi, nilai minimum, dan nilai maksimum. Pada penerapan ini peubah curah hujan diperlihatkan karena memiliki perubahan lebih banyak dibandingkan pada peubah lainnya.

TABEL 10
PERBANDINGAN STATISTIK DESKRIPTIF PEUBAH CURAH HUJAN

Ukuran Statistik	Nilai Sebelum Imputasi	Nilai Sesudah Imputasi
Mean	10,8	10,7
Median	1,1	2,5
Deviasi Standar	23,5	22,2
Nilai Minimum	0,0	0,0
Nilai Maksimum	608,2	608,2

Tabel 10 memperlihatkan bahwa metode *mean* per bulan tidak memberikan perubahan signifikan terhadap distribusi data curah hujan. Nilai *mean* dan standar deviasi sebelum dan sesudah imputasi relatif sama. Hal tersebut menunjukkan bahwa mean per bulan tidak mengubah kecenderungan rata-rata dan keragaman data curah hujan secara signifikan. Nilai median memang mengalami sedikit peningkatan dari 1,1 menjadi 2,5, namun hal ini masih bisa diterima karena distribusi data curah hujan yang cenderung menyerupai gamma akibat beberapa kejadian ekstrem [31] dan dominasi nilai rendah. Selain itu, nilai minimum dan maksimum tidak berubah, sehingga informasi mengenai kejadian ekstrem tetap terjaga. Secara keseluruhan, metode *mean* per bulan mampu mempertahankan karakteristik utama distribusi data curah hujan pada Data GSOD.

Hasil ini mengindikasikan bahwa metode sederhana merupakan alternatif yang efisien dan andal untuk mengolah data curah hujan GSOD di wilayah tropis. Kestabilan distribusi data setelah di imputasi sangat penting karena mendukung penerapan langsung pada model hidrologi seperti perhitungan debit puncak dan simulasi banjir. Selain itu, metode ini juga dapat diterapkan pada analisis multivariat seperti PCA dan klasterisasi. Kemampuan metode *mean* per bulan untuk mempertahankan struktur multiariat data sangat penting karena umumnya digunakan untuk mengidentifikasika pola variabilitas iklim [32] dan mereduksi dimensi sebelum menganalisis [33]. Dengan demikian, metode ini memastikan bahwa analisis PCA lanjutan akan menghasilkan komponen utama yang memperkecil bias dan merepresentasikan hubungan iklim yang sesungguhnya.

IV. KESIMPULAN

Berdasarkan hasil penelitian terhadap empat metode imputasi data hilang multivariat pada data iklim Kota Balikpapan dengan proporsi data hilang 11,4%, metode *mean* per bulan terbukti unggul karena stabil dan mendekati struktur data aslinya dengan *relative difference* 2,64%. Metode ini sesuai untuk data dengan pola musiman dan, secara umum, lebih tepat diterapkan pada mekanisme MCAR. Keunggulan metode ini didukung oleh karakteristik iklim tropis Balikpapan yang memiliki pola musiman stabil. Sementara itu, *missRanger* dan k-NN berjalan cukup baik pada kedua skenario, namun evaluasi visual memperlihatkan bahwa struktur hasil imputasinya tidak sebaik *mean* per bulan. Metode ini fleksibel untuk digunakan pada mekanisme MCAR, MAR, maupun MNAR. Adapun *Iterative Robust-Model Imputation* (IRMI), meskipun secara teori dirancang untuk menangani data multivariat dengan MAR, menunjukkan performa terburuk dalam penelitian ini. IRMI gagal dijalankan pada skenario awal dan memberikan hasil prediksi paling buruk pada skenario kedua. Penerapan metode *mean* per bulan pada Data GSOD juga menunjukkan bahwa distribusi data tetap terjaga, sehingga mendukung validitas untuk analisis klimatologi lanjutan seperti perhitungan debit puncak, PCA, dan klasterisasi. Temuan ini menegaskan pentingnya identifikasi awal pola data hilang dan pemilihan metode imputasi yang sesuai dengan karakteristik data,

mekanisme kehilangan data, serta tujuan analisis untuk menjamin kualitas informasi bagi pengambilan kebijakan lingkungan.

DAFTAR PUSTAKA

[1] S. R. Wicaksono, *Prinsip Dasar Kualitas Data*. Malang: Seribu Bintang, 2023. doi: 10.5281/zenodo.12155308.

[2] F. Rafii and T. Kechadi, "Collection of historical weather data: Issues with missing values," *ACM Int. Conf. Proceeding Ser.*, no. 365, 2019, doi: 10.1145/3368756.3368974.

[3] A. Little and B. Rubin, *Analysis with missing*, 3rd ed. Hokoben: Wiley, 2020.

[4] G. Gunawan, "Analisis data hidrologi sungai air bengkulu menggunakan metode statistik," *J. Inersia*, vol. 9, no. 1, pp. 47–58, 2017.

[5] C. Ocampo-marulanda *et al.*, "Missing data estimation in extreme rainfall indices for the Metropolitan area of Cali - Colombia : An approach based on artificial neural networks," *Data Br.*, vol. 39, p. 107592, 2021, doi: 10.1016/j.dib.2021.107592.

[6] B. Gomer, "Mcar, mar, and mnar values in the same dataset: a realistic evaluation of methods for handling missing data," *Multivariate Behav. Res.*, vol. 54, no. 1, p. 153, 2019, doi: 10.1080/00273171.2018.1557033.

[7] P. Saeipourdizaj, P. Sarbakhsh, and A. Gholampour, "Application of imputation methods for missing values of pm10 and o3 data: interpolation, moving average and k-nearest neighbor methods," *Environ. Heal. Eng. Manag.*, vol. 8, no. 3, pp. 215–226, 2021, doi: 10.34172/EHEM.2021.25.

[8] M. Templ and M. Ulmer, "The impact of misclassifications and outliers on imputation methods," *J. Appl. Stat.*, vol. 51, no. 14, pp. 2894–2928, 2024, doi: 10.1080/02664763.2024.2325969.

[9] K. Gurtskaia, J. Schwerter, and P. Doebler, "Adapting tree-based multiple imputation methods for multi-level data ? A simulation study," *arXiv Prepr.*, vol., no., p., 2024, doi: 10.48550.

[10] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Diego: Morgan Kaupmann, 2001.

[11] M. Mayer, "Package ' missRanger ,'" pp. 1–10, 2024, doi: 10.1093/bioinformatics/btr597>.

[12] M. N. Wright and A. Ziegler, "Ranger: A fast implementation of random forests for high dimensional data in C++ and R," *J. Stat. Softw.*, vol. 77, no. 1, pp. 1–17, 2017, doi: 10.18637/jss.v077.i01.

[13] J. Schwerter, K. Gurtskaia, A. Romero, B. Zeyer-Gliozzo, and M. Pauly, "Evaluating tree-based imputation methods as an alternative to MICE PMM for drawing inference in empirical studies," *arXiv Prepr.*, vol., p., 2024.

[14] Y. S. Resheff and D. Weinshall, "Optimized linear imputation," in *6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, Setúbal: SCITEPRESS, 2017, pp. 17–25. doi: 10.5220/0006092900170025.

[15] M. Templ, "Enhancing precision in large-scale data analysis: an innovative robust imputation algorithm for managing outliers and missing values," *Mathematics*, vol. 11, no. 12, 2023, doi: 10.3390/math11122729.

[16] M. Templ, A. Kowarik, and P. Filzmoser, "Iterative stepwise regression imputation using standard and robust methods," *Comput. Stat. Data Anal.*, vol. 55, no. 10, pp. 2793–2806, 2011, doi: 10.1016/j.csda.2011.04.012.

[17] C. Li, "Little's test of missing completely at random," *Stata J.*, vol. 13, no. 4, pp. 795–809, 2013, doi: 10.1177/1536867x1301300407.

[18] M. W. Heymans and J. W. R. Twisk, "Handling missing data in clinical research," *J. Clin. Epidemiol.*, vol. 151, pp. 185–188, 2022, doi: 10.1016/j.jclinepi.2022.08.016.

[19] M. Jamshidian and S. Jalal, "Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data," *Psychometrika*, vol. 75, no. 4, pp. 649–674, 2010, doi: 10.1007/s11336-010-9175-3.

[20] P. McKnight, K. McKnight, S. Sidani, and A. Figueredo, *Missing Data: A Gentle Introduction*. New York City: The Guilford Press, 2007.

[21] C. K. Enders, *Missing Applied Analysis Data*. New York City: The Guilford Press, 2010.

[22] M. Martinkova, "Overview of observed clausius-clapeyron scaling of extreme precipitation in midlatitudes," *Atmosphere (Basel).*, vol. 11, pp. 1–16, 2020, doi: 10.3390/atmos11080786.

[23] C. Xu *et al.*, "Asian-Australian summer monsoons linkage to ENSO strengthened by global warming," *npj Clim. Atmos. Sci.*, vol. 6, no. 1, 2023, doi: 10.1038/s41612-023-00341-2.

[24] H. Aguilera, C. Guardiola-Albert, and C. Serrano-Hidalgo, "Estimating extremely large amounts of missing precipitation data," *J. Hydroinformatics*, vol. 22, no. 3, pp. 578–592, 2020, doi: 10.2166/hydro.2020.127.

[25] Y. Zhou, S. Aryal, and M. R. Bouadjenek, "A comprehensive review of handling missing data: exploring special missing mechanisms," 2024.

[26] A. J. Mason, R. D. Grieve, A. Richards-belle, P. R. Mouncey, D. A. Harrison, and J. R. Carpenter, "Open Access A framework for extending trial design to facilitate missing data sensitivity analyses," *BMC Med. Res. Methodol.*, vol. 2, pp. 1–12, 2020, doi: 10.1186/s12874-020-00930-2.

[27] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-nearest neighbor (K-NN) based missing data imputation," in *5th International Conference on Science in Information Technology (ICSITech)*, 2019, pp. 83–88. doi: https://doi.org/10.1109/icsitech46713.2019.8987530.

[28] N. Umar and A. Gray, "Optimal parameter choice for imputing missing values in water level data using the k-nearest neighbour (kNN) method.," in *The Doctoral School Multidisciplinary Symposium (DSMS 2023)*, Glasglow, United Kingdom, 2023, pp. 1–2.

[29] H. Manlea, *Klimatologi Dasar*. Jakarta: PT Literasi Nusantara Abadi Group, 2020.

[30] [BMKG], "Indonesia Typical Meteorogical Year," Badan Meteorologi, Klimatologi, dan Geofisika. Accessed: Nov. 24, 2025. [Online]. Available: https://iklim.bmkg.go.id/id/i-tmy/

[31] C. Martinez-Villalobos and J. D. Neelin, "Why Do Precipitation Intensities Tend to Follow Gamma Distributions ?," *J. Atmos. Sci.*, vol. 76, no. 1, pp. 3611–3631, 2019, doi: 10.1175/JAS-D-18-0343.1.

[32] C. Guilloteau, A. Mamalakis, L. Vulis, P. V. V. Le, T. T. Georgiou, and E. Foufoula-Georgiou, "Rotated spectral principal component cnalysis (rsPCA) for identifying dynamical codes of variability in climate cystems," *J. Clim.*, vol. 34, pp. 715–736, 2021, doi: 10.1175/JCLI-D-20-0266.1.

[33] G. Sottile, A. Francipane, G. Adelfino, and L. V. Noto, "A PCA-based clustering algorithm for the identification of stratiform and convective precipitation at the event scale : an application to the sub-hourly precipitation of Sicily , Italy," *Stoch. Environ. Res. Risk Assess.*, vol. 36, no. 8, pp. 2303–2317, 2022, doi: 10.1007/s00477-021-02028-7.