Megasains, Vol. 14, No. 2 , 23-31
p-ISSN 2086-5589 e-ISSN : 2723-2239
https://megasains.gawbkt.id

megasains
buletin meteorologi, klimatologi, kualitas udara, geofisika, dan lingkungan

PENERAPAN IMPUTASI LOCF DAN *CROSS MEAN* DALAM PENGISIAN DATA KOSONG PADA CURAH HUJAN HARIAN ARG

APPLICATION OF LOCF AND CROSS MEAN IMPUTATION IN COMPLETING MISSING DATA ON ARG DAILY RAINFALL

Siti Risnayah[1]*, La Ode Hasnuddin S. Sagala[2]
[1]BMKG, Stasiun Klimatologi Sulawesi Tenggara, Jl. Poros Bandara Kec. Ranomeeto, Konawe Selatan, 93870
[2]Fakultas Teknologi Informasi, Universitas Sembilanbelas November Kolaka, Kec. Tanggetada, Kolaka, 93561
**E-mail: siti.risnayah@bmkg.go.id*

Naskah masuk: 28 September 2023 Naskah diperbaiki: 20 Oktober 2023 Naskah diterima: 31 Oktober 2023

ABSTRAK

Banyaknya alat penakar hujan Automatic Rain Gauge (ARG) yang telah terpasang saat ini belum termanfaatkan secara optimal. Hal ini disebabkan ARG yang bekerja secara otomatis sering kali mengalami *missing data* akibat permasalahan teknis dan jaringan yang pada akhirnya menimbulkan keraguan akan keakuratannya. Data yang digunakan dalam penelitian ini adalah data curah hujan ARG dalam periode 10 menit selama tahun 2021 serta data curah hujan dari alat penakar hujan konvensional di lokasi yang sama. Data akan diolah hingga menjadi data harian kemudian dipulihkan dengan pengisian data kosong yang dikerjakan melalui bahasa pemrograman Python. Karena data ARG berjenis data longitudinal maka pengisian data kosong akan menggunakan imputasi LOCF dan *cross mean*. Uji validitas kemudian dilakukan untuk membandingkan data ARG yang telah dipulihkan dengan data dari alat manualnya melalui perhitungan nilai MAE, RMSE, dan koefisien korelasi. Hasil penelitian menunjukkan bahwa pengisian data kosong dapat mengurangi persentase *missing data* yang semula 21.4% menjadi 1.1%. Hasil uji validitas menunjukkan ARG dapat menghasilkan data yang akurat ditandai dengan nilai eror yang rendah (MAE=0.998 mm, RMSE=2.253 mm) dan korelasi yang sangat tinggi (r=0.966). Dengan semakin tingginya persentase kelengkapan data dan akurasi yang sangat baik maka penggunaan data tersebut akan menjadi semakin luas sehingga memberikan lebih banyak manfaat terutama untuk keperluan analisis, prakiraan, pelayanan data, maupun penelitian.

Kata kunci: ARG, *missing data*, Python, LOCF, *Cross Mean*

ABSTRACT

The number of installed Automatic Rain Gauges (ARG) today has not been optimally utilized. It is because ARG that works automatically often has missing data due to technical and network problems raising doubts about its accuracy. The data used are ARG rainfall data in 10 minute periods during 2021 and rainfall data from conventional gauge at the same location. The data will be processed until it becomes daily data and will be recovered by missing data entry worked by the Python programming language. Because the ARG data is the longitudinal data type, missing data entry will use LOCF and cross mean imputation. The validity test will compare the recovered ARG data with the conventional gauge data by calculating the MAE, RMSE, and correlation coefficient. The results showed that missing data entry could reduce the percentage of missing from 21.4% to 1.1%. The result of validity tests showed that ARG could produce accurate data determined by a lower error (MAE=0.998mm, RMSE=2.253mm) and a very high correlation (r=0.966). With a higher percentage of data completeness and excellent accuracy, the data usage will become more extensive to provide more benefits, especially for the need of analysis, forecasting, data services, and research.

***Keywords*:** *ARG, missing data, Python, LOCF, Cross Mean*

DOI : 10.46824/megasains.v14i2.138

1. Pendahuluan

Pada bidang klimatologi, bahasa pemrograman Python sangat handal digunakan untuk pengolahan data numerik seperti data *timeseries* curah hujan keluaran *Automatic Rain Gauge* (ARG). ARG merupakan alat penakar hujan otomatis yang mencatat curah hujan setiap 10 menit dan menyimpannya dalam *data logger*. Permasalahan yang sering kali terjadi pada setiap alat otomatis adalah adanya *missing data* akibat kendala jaringan ataupun kendala teknis lainnya. *Missing data* yang terjadi terkadang masih dapat diselamatkan karena beberapa ARG mengakumulasikan datanya di menit berikutnya. Jika tidak diamati dengan baik maka *missing data* akan dibuang padahal data tersebut masih dapat dipulihkan.

Terdapat dua format data curah hujan pada ARG yakni yang mengakumulasikan hujan ke menit berikutnya dan yang menampilkan data hujan *real* setiap 10 menit. Pada ARG yang mengakumulasi curah hujannya hingga jam 00.00 UTC, total curah hujan hariannya adalah data pada jam 00.00 UTC tersebut. Akan tetapi dalam penentuannya sebagai total curah hujan harian klimatologi di BMKG didasarkan pada laporan hujan F-Klim71 yang juga diterapkan di pos hujan kerjasama yakni jumlah curah hujan harian dihitung dari jam 07.00 hingga jam 07.00 waktu setempat keesokan harinya dan dicatat sebagai hujan pada waktu penakaran [1]. Mengingat wilayah penelitian masuk dalam zona Waktu Indonesia Tengah (WITA), maka jam 07.00 waktu setempat adalah jam 23.00 UTC. Oleh sebab itu, curah hujan harian ARG diperoleh dari curah hujan yang tercatat pada jam 23.00 UTC ditambah dengan selisih antara jam 00.00 dan 23.00 UTC hari sebelumnya.

Banyaknya alat otomatis yang telah dipasang saat ini belum menjadi solusi ketersediaan data iklim di banyak lokasi di Indonesia. Hal ini disebabkan karena pengolahan data dari alat otomatis ini belum dilakukan secara optimal. Kesulitan memahami struktur data, kondisi data yang masih mentah, dan terutama karena kurangnya kepercayaan akan data dari alat otomatis menjadi faktor penyebabnya [2]. Karena data terkumpul secara otomatis, ada ketakutan bahwa data tidak akurat mengingat banyaknya permasalahan yang sering dialami oleh alat otomatis. Alat otomatis yang jarang dilakukan pengecekkan seringkali data tetap masuk akan tetapi nilainya menunjukkan angka yang ‘aneh’ sehingga performa alat ini pun semakin diragukan.

Tahap awal dalam pengolahan data adalah proses pembersihan data (*data cleaning*). Salah satu langkah dalam pembersihan data ini adalah proses pembersihan data kosong yang mana dapat ditangani dengan menghapusnya atau mengisi nilainya dengan estimasi atau imputasi tergantung kondisi data tersebut [3]. Pengisian data kosong bertujuan untuk memulihkan data sehingga jumlah data semakin banyak dan pengolahannya semakin akurat. Dampak dari data yang hilang pada penelitian kuantitatif dapat menjadi serius karena dapat menyebabkan perkiraan parameter yang bias, hilangnya informasi, penurunan kekuatan statistik, peningkatan kesalahan standar, dan generalisasi temuan yang melemah [4]. Persentase data masuk yang rendah secara tidak langsung dapat mengurangi tingkat kepercayaan data sehingga data tidak digunakan dalam analisis [5, 6].

Pemulihan data akan sangat berguna bagi kualitas data itu sendiri. Kondisi data yang cenderung lengkap dapat meningkatkan daya guna data tersebut sehingga dapat dimanfaatkan dan digunakan dalam analisis, prakiraan, maupun penelitian. Tidak ada batasan yang pasti mengenai persentase yang dapat diterima dari data yang hilang untuk kesimpulan statistik yang valid [4]. Terdapat referensi yang menyatakan ambang maksimum yang aman untuk data kosong adalah 5% dari total kumpulan data yang besar, jika lebih dari 5% maka sebaiknya tidak digunakan [6]. Beberapa referensi menetapkan aturan berbeda untuk persentase *missing data* yang berbeda [7, 8]. Scheffer (2002) menyarankan penghapusan *missing data* dapat digunakan jika tidak lebih dari 6% data hilang, imputasi tunggal jika tidak lebih dari 10% data hilang dan prosedur yang lebih kompleks seperti imputasi berganda jika antara 10% dan 25% dari datanya hilang [7]. Zhang dkk. (2004) dalam Ocampo-Marulanda dkk (2021) menetapkan curah hujan bulanan tidak dihitung jika terdapat *missing data* lebih dari tiga hari dalam sebulan dan curah hujan tahunan tidak dihitung jika lebih dari 15 hari atau satu bulan hilang dalam setahun. [9].

World Meteorological Organization (WMO) sebagai badan dunia yang membuat aturan pengolahan data iklim bahkan menetapkan aturan yang lebih ketat. Dalam perhitungan total curah hujan bulanan, WMO mensyaratkan tidak boleh ada satu haripun data yang hilang [10]. WMO mensyaratkan total bulanan agar dapat digunakan harus memiliki kelengkapan 100% baik data itu dibangun dari hasil pengamatan maupun estimasi [10].

p-ISSN: 2086-5589
e-ISSN: 2723-2239

Diterangkan lebih lanjut bahwa data estimasi harus memenuhi kriteria 11/5 yakni menerima hanya 11 hari data estimasi dan tidak lebih 5 hari berturut-turut [10]. Hal ini tentu menjadi kendala bagi kebanyakan alat otomatis untuk memenuhi syarat penggunaan data karena seringkali mengalami gangguan jaringan yang kemudian menghasilkan *missing value*. Melalui proses pengisian data kosong yang tepat maka penggunaan data meningkat dan informasi yang dihasilkan tidak menjadi bias.

Saat ini pemikiran bagaimana melakukan pengisian *missing value* menjadi lebih berkembang daripada pemikiran bagaimana mengatasi *missing value* tersebut [11]. Telah banyak penelitian terkait pengisian data-data yang hilang dari rangkaian curah hujan dimana hasil akhirnya dapat meningkatkan akurasi data tersebut. Miro dkk (2017) telah mengujicobakan beberapa pendekatan untuk mengisi seris data yang tidak lengkap dan berhasil menemukan satu metode yang paling baik [11]. Muflihah dan Pahlawan (2017) juga menyatakan dapat meningkatkan korelasi dan mengurangi RMSE dari pengisian data kosong curah hujan dengan metode yang tepat [12]. Ocampo-Marulanda dkk (2021) berhasil melakukan pengisian data kosong pada kasus hujan ekstrim ditandai dengan koefisien korelasi mendekati 1 dan RMSE mendekati 0 [9]. Sementara Papailiou dkk (2022) menemukan metode yang lebih akurat untuk mengisi data hujan yang hilang akan tetapi membutuhkan waktu yang lebih lama [13]. Penelitian ini juga akan melakukan pengolahan data *timeseries* curah hujan yang dilakukan secara otomatis menggunakan bahasa pemrograman Python dimana dalam prosesnya akan dilakukan pemulihan data yang hilang. Diharapkan penelitian ini akan memudahkan pengolahan data ARG sehingga ke depannya data dari alat otomatis ini dapat digunakan dalam analisis, prakiraan, pelayanan data, maupun penelitian.

2. Metode Penelitian

Data yang digunakan dalam penelitian ini adalah data curah hujan hasil pengukuran alat otomatis jenis *Automatic Rain Gauge* (ARG) dalam periode 10 menit selama tahun 2021 serta data curah hujan harian hasil observasi dengan alat penakar hujan konvensional tipe OBS di lokasi yang sama (lihat Tabel 1). Data alat otomatis ARG diperoleh dari situs *awscenter.bmkg.go.id* sementara data pengukuran manual diperoleh dari BMKGSoft hasil input data elektronik Pos Hujan Kerjasama (ePHK).

Tabel 1. Metadata ARG dan Pos Hujan Kerjasama (PHK)

Parameter	ARG	PHK
Nama	ARG Kapontori	PHK Wakangka
ID Stasiun	150159	74042201a
Lintang	5.20872 LS	5.20900 LS
Bujur	122.8283 BT	122.8280 BT
Lokasi	Desa Wakangka, Kec. Kapontori, Kab. Buton	Desa Wakangka, Kec. Kapontori, Kab. Buton
Reset Time	00.10 UTC (08.10 WITA)	23.00 UTC (07.00 WITA)

Sumber: BMKG

Data curah hujan yang telah terkumpul akan diolah menggunakan bahasa pemrograman Python yang dikerjakan melalui *Google Colabs*. Sebuah *script* sederhana akan disusun sehingga nantinya petugas pengolah data cukup dengan memasukkan data unduhan dari ARG dan mengisi beberapa informasi lalu me-*running script*-nya sehingga menghasilkan data harian. Dalam *script* tersebut telah mencakup pengisian data kosong sehingga jumlah data yang dihasilkan dapat maksimal. Selanjutnya data akan divisualisasikan dan dibandingkan dengan data dari alat konvensional untuk diverifikasi. Urutan pengolahan data ARG menggunakan Python dijelaskan dalam poin-poin di bawah ini:

1) Unduh data curah hujan ARG dan pos hujan dalam format Excel dan simpan di komputer atau *Google Drive*.
2) *Rename file* unduhan tersebut dengan format arg_stasiun_tahun_bulan. Contoh: arg_kapontori_2021_01.
3) Sedikit modifikasi pada data ARG yakni menghapus keseluruhan baris pertama yang berisi keterangan *Quality Code*.
4) Menggunggah data ARG dan data pos hujan tersebut ke dalam Python.
5) Cukup mengisi 3 informasi yakni list periode data yang diolah, nama stasiun, dan list periode data yang akan dihasilkan.

Jenis data ARG yang mengakumulasikan hujan ke menit berikutnya hingga ke jam 00.00 UTC (dan kembali 0 mm pada jam 00.10 UTC) adalah jenis data longitudinal. Terdapat banyak metode imputasi dalam mengatasi kehilangan data pada data longitudinal, namun metode yang paling cocok digunakan adalah imputasi tunggal berupa LOCF dan *cross mean* [7, 14, 15]. LOCF (*Last Observation Carried Forward*) berarti mengganti nilai hilang dengan nilai pengamatan terakhir yang dapat tercatat sementara *cross mean* berarti mengganti nilai hilang dengan nilai rata-rata [5, 14, 15]. Pengisian data kosong dengan nilai pengamatan terakhir sangat berguna pada data

p-ISSN: 2086-5589
e-ISSN: 2723-2239

dimana nilai pengamatan akan dicatat hanya ketika ada perubahan [14].

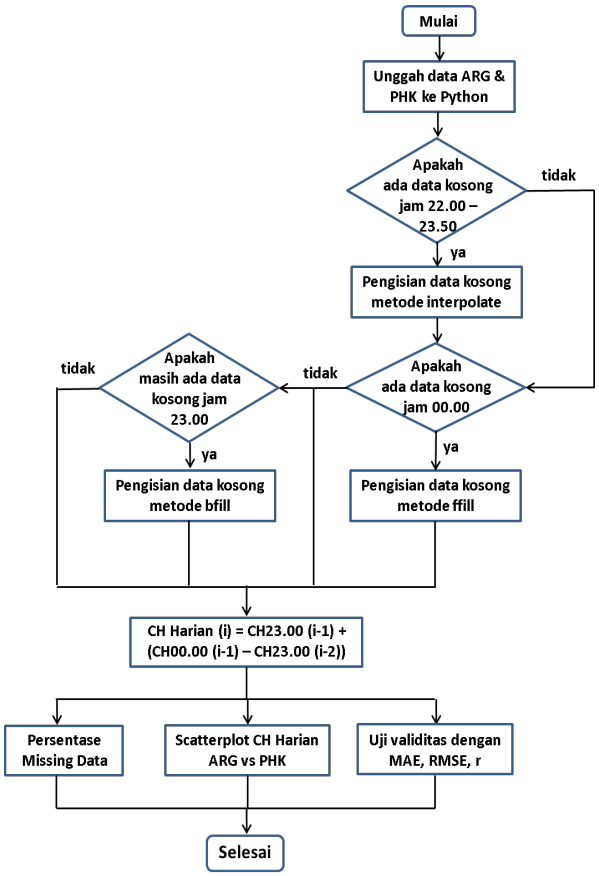
Waktu utama untuk menentukan total curah hujan harian pada data ARG adalah jam 23.00 dan 00.00 UTC. Untuk memulihkan nilai-nilai yang tidak terekam pada waktu tersebut maka pada penelitian ini ditetapkan batasan waktu yang menjadi pengisi data hilang tersebut yakni data antara jam 22.00 s.d 00.00 UTC. Batasan ini ditetapkan dengan keyakinan bahwa curah hujan 1 jam sebelum dan sesudah jam utama (22.00 UTC dan 00.00 UTC) dapat sama atau minimal mendekati dengan hujan pada jam utama (23.00 UTC).

Untuk lebih memahami teknis pengisian data kosong, berikut dijelaskan jenis-jenis data kosong yang terdapat pada data ARG.

1) Data kosong sebagian antara jam 22.00 – 00.00 UTC maka data masih bisa dipulihkan dengan nilai pada rentang jam tersebut.
2) Data kosong keseluruhan dari jam 22.00 – 00.00 UTC maka data tidak bisa dipulihkan dan akan tercatat sebagai NaN atau 9999.

Pada bahasa pemrograman Python, terdapat sintaks untuk imputasi tunggal yakni *interpolate, forward fill* (*ffill*), dan *backward fill* (*bfill*). Sintaks-sintaks ini akan dipakai dalam 3 tahapan seperti yang dijelaskan pada diagram alir (Gambar 1). Sintaks *interpolate* digunakan pada pencarian nilai tengah dari 2 data yang ada, sintaks *ffill* digunakan pada pengisian data kosong dengan data dari menit sebelumnya, sedangkan sintaks *bfill* digunakan pada pengisian data kosong dengan data dari menit setelahnya [16]. Dengan prinsip rata-rata, imputasi *interpolate* mampu mengisi data kosong yang diapit oleh dua data antara jam 22.00 – 23.50 UTC sehingga data jam utama 23.00 UTC dapat dipulihkan (tidak dapat dipulihkan jika tidak ada data selama jam 22.00 – 23.50 UTC). Selanjutnya mengecek jam 00.00 UTC, jika kosong maka dilengkapi dengan data dari jam 23.50 UTC berdasarkan prinsip *ffill*. Kemudian kembali mengecek kelengkapan jam 23.00 UTC, jika masih kosong maka dilengkapi dengan data dari jam 23.50 UTC berdasarkan prinsip *bfill.* Untuk kasus *missing data* seharian dan *missing data* di atas jam 22.00 UTC maka data dibiarkan hilang atau ditulis sebagai NaN.

Menggunakan imputasi tunggal dapat menghasilkan nilai pengisian yang terlalu rendah [16] terutama ketika kasus kejadian hujan lebat antara jam 23.00 – 00.00 UTC tidak terekam. Untuk meningkatkan keyakinan pada metode ini diperlukan analisis statistik tambahan



Gambar 1. Diagram Alir Pengolahan Data ARG (Sumber: data diolah)

yang tepat yang dapat membedakan antara nilai sebenarnya dan nilai estimasi [14]. Dalam penelitian ini akan digunakan perhitungan statistik uji validitas dengan menghitung nilai MAE, RMSE, koefisien korelasi, dan *slope* menggunakan *library metrics* dan *pandas* yang tersedia di Python. Nilai MAE dan RMSE dikatakan baik jika nilainya mendekati 0 sedangkan koefisien korelasi baik jika besarannya mendekati 1 [17]. *Slope* berarti rata-rata pertambahan (jika bernilai positif) atau pengurangan (jika bernilai negatif) yang terjadi pada variabel *y* untuk setiap peningkatan satu satuan variabel *x* [18]. Nilai statistik ini dapat menjadi koreksi maupun verifikasi performa alat maupun kualitas kedua data tersebut.

3. Hasil dan Pembahasan

Tipe Data Kosong pada ARG. Untuk memahami bagaimana pengisian data kosong dapat dilakukan, pertama-tama perlu mengetahui kondisi data mentah ARG yang seringkali kehilangan data. Beberapa kejadian data hilang di ARG Kapontori yang sebagian datanya masih dapat diselamatkan ditampilkan pada Gambar 2. Mengingat jam utama dalam

p-ISSN: 2086-5589
e-ISSN: 2723-2239

Time	1/21	1/22	2/3	2/4	3/10	3/18	6/13	6/21	6/26	6/29	7/1	7/2	7/3	7/5	7/15	7/16	7/26	7/27	7/28	8/10	8/11	9/23	9/24
0:00:00	3		7.4		1.4	0	10.6	37.6	2.8	18.6	0.8	3.2	30.6	0.6	5.6	7.4	0	0	0	0.2	1.8	18	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
22:00:00		7.6	12	19.2	13	7.2	1	66.6	31.6						6.6	0				1.2	3.2		0.2
22:10:00		7.6	12.2	19.2	13	7.2	1	66.6							6.6	0				1.2	3.2		0.2
22:20:00		7.6		19.2	13	7.2	1	66.6	31.6						6.8	0				1.2	3.2		0.2
22:30:00		7.6		19.2	13		1	66.6	31.6	21.6	3.2	30.4	4.6		7	0				1.2	3.2		0.2
22:40:00		7.6		19.2		7.2	1	66.6	31.8						7	0				1.4	3.2		0.2
22:50:00		7.6		19.2	13	7.2	1	66.6	31.8						7	0				1.4	3.2		0.2
23:00:00		7.6		19.2		7.2	1	66.6	31.8						7.2	0				1.4	3.2		0.2
23:10:00		7.6		19.2	13	7.2	1	66.6	31.8						7.4	0				1.4	3.2		0.2
23:20:00		7.6		19.2			1	66.6	31.8							0				1.4	3.2		0.2
23:30:00		7.6		19.2	13		1		31.8		3.2	30.6				0							0.2
23:40:00		7.6		19.2		7.2	1	66.6								0							0.2
23:50:00		7.6		19.2	13	7.2		66.6	32	21.6	3.2		4.6	1.4		0							0.2
	x	x	v	v	v	v	v	v	v	v	v	v	v	v	v		v	v		v	v	x	x

Gambar 2. Sampel Data Curah Hujan di ARG Kapontori dalam Milimeter; Baris Atas Bulan/Tanggal; Baris Bawah X Tidak Dapat Dipulihkan dan V Dapat Dipulihkan; Merah Muda Berarti Data Tidak Ada
(Sumber: data diolah)

menentukan akumulasi hujan harian pada Kecamatan Kapontori (Zona waktu WITA) adalah jam 00.00 dan 23.00 UTC maka data harus tersedia pada waktu tersebut. Gambar 2 telah mengkategorikan hari-hari yang datanya bisa dipulihkan dan yang tidak bisa. Baris berwarna biru yang ditandai dengan tanda silang 'X' yakni tanggal 21 Januari jam 23.00, 22 Januari jam 00.00, 23 September jam 23.00 dan 24 September jam 00.00 tidak dapat dipulihkan datanya karena tidak ada data sama sekali sepanjang jam 22.00 hingga 00.00 UTC sehingga data pada hari tersebut dianggap hilang atau 9999. Sebaliknya pada kolom yang diberi tanda ceklis 'V' berarti datanya masih dapat dipulihkan. Contohnya *missing data* jam 23.00 UTC pada tanggal 3 Februari dapat diisi dengan nilai 12.2 mm, tanggal 10 Maret 13.0 mm, tanggal 29 Juni 21.6 mm, tanggal 2 Juli 30.5 mm, 5 Juli 1.4 mm, 26 Juli 0.0 mm, dan 27 Juli 0.0 mm.

Pengisian Data Kosong Tahap I. Tahap awal pengisian data kosong adalah pengisian pada jam 23.00 UTC berdasarkan data hujan antara jam 22.00 s.d 23.50 UTC menggunakan sintaks *interpolate*. Beberapa data dapat dipulihkan seperti yang ditunjukkan oleh Gambar 3 pada kolom yang berwarna hijau. Dapat dilihat nilai pemulihan pada Gambar 3 sama dengan nilai dugaan seperti pada penjelasan sebelumnya (analisis pada Gambar 2).

Time	1/21	1/22	2/3	2/4	3/10	3/18	6/13	6/21	6/26	6/29	7/1	7/2	7/3	7/5	7/15	7/16	7/26	7/27	7/28	8/10	8/11	9/23	9/24
22:00:00		7.6	12	19.2	13	7.2	1	66.6	31.6						6.6	0				1.2	3.2		0.2
22:10:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6						6.6	0				1.2	3.2		0.2
22:20:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6						6.8	0				1.2	3.2		0.2
22:30:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6	21.6	3.2	30.4	4.6		7	0				1.2	3.2		0.2
22:40:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.4	4.6		7	0				1.4	3.2		0.2
22:50:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6		7	0				1.4	3.2		0.2
23:00:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6		7.2	0				1.4	3.2		0.2
23:10:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6		7.4	0				1.4	3.2		0.2
23:20:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:30:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:40:00		7.6	12.2	19.2	13	7.2	1	66.6	31.9	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:50:00		7.6	12.2	19.2	13	7.2	1	66.6	32	21.6	3.2	30.6	4.6	1.4	7.4	0				1.4	3.2		0.2

Gambar 3. Sampel Data Curah Hujan yang Telah Mengalami Pengisian Data Kosong Tahap I (Merah Muda: Data Belum Dapat Dipulihkan, Hijau: Data Hasil Pemulihan)
(Sumber: data diolah)

Time	1/21	1/22	2/3	2/4	3/10	3/18	6/13	6/21	6/26	6/29	7/1	7/2	7/3	7/5	7/15	7/16	7/26	7/27	7/28	8/10	8/11	9/23	9/24
0:00:00	3		7.4	12.2	1.4	0	10.6	37.6	2.8	18.6	0.8	3.2	30.6	0.6	5.6	7.4	0	0	0	0.2	1.8	18	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
22:00:00		7.6	12	19.2	13	7.2	1	66.6	31.6						6.6	0				1.2	3.2		0.2
22:10:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6						6.6	0				1.2	3.2		0.2
22:20:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6						6.8	0				1.2	3.2		0.2
22:30:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6	21.6	3.2	30.4	4.6		7	0				1.2	3.2		0.2
22:40:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.4	4.6		7	0				1.4	3.2		0.2
22:50:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6		7	0				1.4	3.2		0.2
23:00:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6		7.2	0				1.4	3.2		0.2
23:10:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6		7.4	0				1.4	3.2		0.2
23:20:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:30:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:40:00		7.6	12.2	19.2	13	7.2	1	66.6	31.9	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:50:00		7.6	12.2	19.2	13	7.2	1	66.6	32	21.6	3.2	30.6	4.6	1.4	7.4	0	0	0	10	1.4	3.2		0.2

Gambar 4. Sampel Data Curah Hujan yang Telah Mengalami Pengisian Data Kosong Tahap II (Merah Muda: Data Belum Dapat Dipulihkan, Hijau: Data Hasil Pemulihan)
(Sumber: data diolah)

p-ISSN: 2086-5589
e-ISSN: 2723-2239

Pengisian Data Kosong Tahap II. Pengisian data kosong tahap II adalah pengisian data kosong pada jam 00.00 UTC berdasarkan data hujan sebelumnya antara jam 22.00 s.d 23.50 UTC menggunakan sintaks *ffill* dan pengisian data kosong pada jam 23.50 UTC berdasarkan data hujan 00.00 UTC hari selanjutnya menggunakan sintaks *bfill*. Beberapa data dapat dipulihkan seperti yang ditunjukkan oleh Gambar 4 pada kolom yang berwarna hijau.

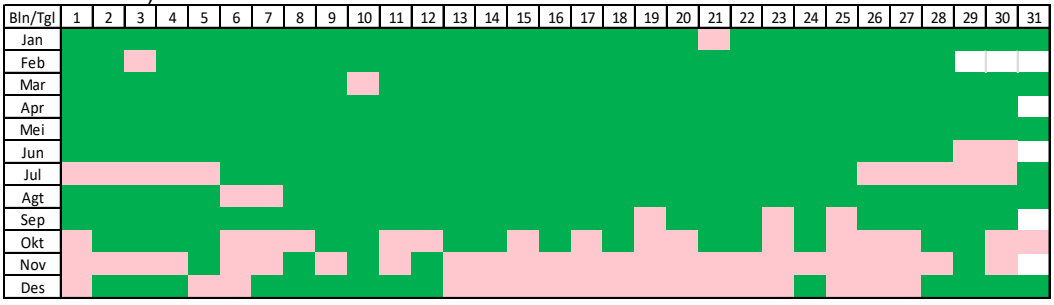
Pengisian Data Kosong Tahap III. Tahap akhir pengisian data kosong adalah pengisian kembali pada jam 23.00 UTC berdasarkan data hujan pada jam 23.50 UTC yang telah dipulihkan pada tahap II sebelumnya menggunakan sintaks *bfill.* Beberapa data dapat dipulihkan seperti yang ditunjukkan oleh Gambar 5 pada kolom yang berwarna hijau.

Analisis Sebelum dan Sesudah Pengisian Data Kosong. Sejauh ini pengisian data kosong telah berhasil dilakukan dan menyisakan beberapa data yang tidak dapat dipulihkan. Gambar 6 menampilkan kondisi data mentah sebelum dilakukannya pengolahan menggunakan Python. Dapat dilihat kondisi data awal menunjukkan ada banyak data yang hilang secara kontinyu. Terutama pada November dan Desember secara berturut-turut data pada jam utama (23.00 UTC) tidak terekam. Jika berdasarkan pada kriteria yang ditetapkan Zhang dkk dalam Ocampo-Marulanda dkk (2021) [9] maka data bulanan Juli, Oktober, November dan Desember tidak akan dihitung sehingga data tahun 2021 pun tidak akan digunakan. Jika berdasarkan aturan WMO [10] maka hanya bulan April dan Mei saja yang dapat digunakan. Hal ini sangat disayangkan mengingat kekosongan data terjadi hanya pada jam penentu yakni 23.00 UTC saja namun masih terdapat data pada rentang waktu 1 jam sebelum dan setelahnya. Setelah dilakukan pengolahan dan pemulihan data, kuantitas data menjadi semakin membaik seperti yang ditampilkan Gambar 7.

Gambar 7 membuktikan data hilang pada jam 23.00 UTC dapat dipulihkan secara maksimal dengan keberadaan data pada rentang 22.00 s.d 00.00 UTC. Dalam perhitungan total harian, kehilangan data di jam 23.00 UTC dalam 1 hari akan menyebabkan kehilangan data curah hujan harian dalam 2 hari. Contoh kasus pada tanggal 21 Januari 2021 (Gambar 6) hilang 1 hari saja akan tetapi setelah dipulihkan (Gambar 7) data *missing* menjadi 2 hari yakni tanggal 22 dan 23 Januari. Hal ini disebabkan oleh total harian dihitung berdasarkan data jam 23.00 UTC ditambah dengan selisih hujan 23.00 UTC dan 00.00 UTC hari sebelumnya dan dicatat di hari pengukuran. Oleh sebab itu dalam mengkonversi hujan ARG dari zona waktu standar UTC menjadi zona waktu lokal WITA ada pergeseran tanggal selama 1 hari.

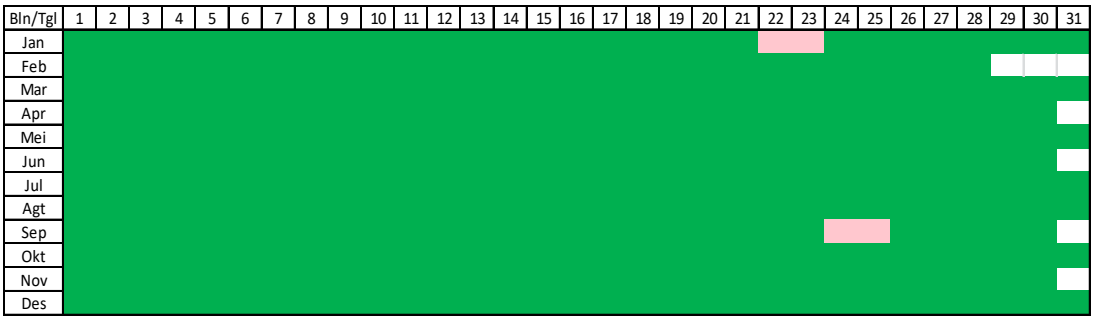
Time	1/21	1/22	2/3	2/4	3/10	3/18	6/13	6/21	6/26	6/29	7/1	7/2	7/3	7/5	7/15	7/16	7/26	7/27	7/28	8/10	8/11	9/23	9/24
0:00:00	3		7.4	12.2	1.4	0	10.6	37.6	2.8	18.6	0.8	3.2	30.6	0.6	5.6	7.4	0	0	0	0.2	1.8	18	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
22:00:00		7.6	12	19.2	13	7.2	1	66.6	31.6						6.6	0				1.2	3.2		0.2
22:10:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6						6.6	0				1.2	3.2		0.2
22:20:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6						6.8	0				1.2	3.2		0.2
22:30:00		7.6	12.2	19.2	13	7.2	1	66.6	31.6	21.6	3.2	30.4	4.6		7	0				1.2	3.2		0.2
22:40:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.4	4.6		7	0				1.4	3.2		0.2
22:50:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6		7	0				1.4	3.2		0.2
23:00:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6	1.4	7.2	0	0	0	10	1.4	3.2		0.2
23:10:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.5	4.6		7.4	0				1.4	3.2		0.2
23:20:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:30:00		7.6	12.2	19.2	13	7.2	1	66.6	31.8	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:40:00		7.6	12.2	19.2	13	7.2	1	66.6	31.9	21.6	3.2	30.6	4.6		7.4	0				1.4	3.2		0.2
23:50:00		7.6	12.2	19.2	13	7.2	1	66.6	32	21.6	3.2	30.6	4.6	1.4	7.4	0	0	0	10	1.4	3.2		0.2

Gambar 5. Sampel Data Curah Hujan yang Telah Mengalami Pengisian Data Kosong Tahap III (Merah Muda: Data Tidak Dapat Dipulihkan, Hijau: Data Hasil Pemulihan)
(Sumber: data diolah)



Gambar 6. Kelengkapan Data Curah Hujan ARG Kapontori Tahun 2021 Hasil Unduhan di web *www.awscenter.bmkg.go.id* (Hijau: Ada Data, Merah Muda: Tidak Ada Data)
(Sumber: data diolah)

p-ISSN: 2086-5589
e-ISSN: 2723-2239



Gambar 7. Kelengkapan Data Curah Hujan ARG Kapontori Tahun 2021 Setelah Dilakukan Pengisian Data Kosong (Hijau: Ada Data, Merah Muda: Tidak Ada Data)
(Sumber: data diolah)

Metode imputasi tunggal walaupun sederhana tetapi sangat efektif digunakan untuk mengisi celah pada rangkaian data curah hujan ARG. Tabel 2 menunjukkan setelah proses imputasi tunggal pada *missing value*, kini hanya tersisa 4 data saja yang kosong atau sebesar 1.1% dari 21.4% *missing value* di awal. Hal ini berarti data ARG dapat dipulihkan hingga 20.3% atau dengan kata lain kelengkapan data curah hujan harian ARG Kapontori tahun 2021 naik menjadi 98.9%. Empat data yang tidak dapat dipulihkan ini memang sama sekali tidak memiliki data sepanjang jam 22.00 hingga 00.00 UTC hari berikutnya sehingga tidak ada data yang dapat digunakan untuk mengestimasi curah hujan hariannya.

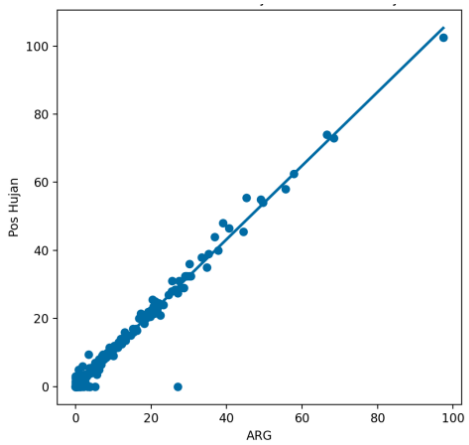
Verifikasi Data Curah Hujan ARG. Untuk menunjukkan keakuratan pengisian data kosong yang dilakukan pada alat penakar hujan otomatis maka perlu dibandingkan dengan data dari alat penakar hujan konvensional. Gambar 8 memvisualisasikan data curah hujan alat otomatis dan konvensional dalam diagram kartesian sumbu x dan sumbu y. Gambar 8 menunjukkan titik-titik data membentuk pola garis lurus dari kiri bawah naik ke kanan atas yang berarti adanya hubungan yang linear dan positif antara kedua data sehingga jika data dari alat otomatis mengalami peningkatan maka data dari alat manual akan meningkat pula. Hal ini sudah sesuai mengingat kedua data adalah variabel yang sama jadi nilainya harus berkolerasi positif.

Tabel 3 menunjukkan keberhasilan proses pengolahan dan pemulihan data curah hujan harian ARG menggunakan bahasa pemrograman Python. Secara statistik, performa ARG dalam mengukur curah hujan di Kapontori sangat baik dengan nilai rata-rata eror (MAE) dan RMSE yang sangat rendah, sebaliknya nilai koefisien korelasi sangat tinggi dimana masing-masing bernilai 0.998 mm, 2.253 mm, dan 0.966. Nilai ini lebih baik dibandingkan sebelum dilakukannya pemulihan data. Dapat dilihat pada Tabel 3 sebelum pemulihan data, nilai MAE dan RMSE lebih tinggi sedangkan nilai koefisien korelasi sedikit

Tabel 2. Statistik Data Kosong Sebelum dan Setelah Dilakukan Pemulihan

Bulan	Sebelum Pemulihan		Setelah Pemulihan	
	Jumlah	%	Jumlah	%
Jan	1	3.2	2	6.5
Feb	1	3.6	-	-
Mar	1	3.2	-	-
Apr	-	-	-	-
Mei	-	-	-	-
Jun	2	6.7	-	-
Jul	10	32.3	-	-
Agt	2	6.5	-	-
Sep	3	10.0	2	6.7
Okt	16	51.6	-	-
Nov	25	83.3	-	-
Des	17	54.8	-	-
Total	78	21.4	4	1.1

Sumber: data diolah



Gambar 8. Diagram *Scatterplot* Curah Hujan ARG vs Pos Hujan Kerjasama di Kapontori
(Sumber: data diolah)

Tabel 3. Hubungan Statistik Data Curah Hujan ARG dan Pos Hujan Kerjasama

Parameter Statistik	Sebelum Pemulihan	Setelah Pemulihan
MAE	1.141 mm	0.999 mm
RMSE	2.548 mm	2.254 mm
r	0.986	0.989
Slope	1.0906	1.0806

Sumber: data diolah

p-ISSN: 2086-5589
e-ISSN: 2723-2239

lebih rendah. Membaiknya akurasi setelah pemulihan sejalan dengan penelitian sebelumnya [12]. Rendahnya nilai eror dan tingginya korelasi mendukung klaim bahwa alat otomatis dapat dengan sangat baik dan memberikan kualitas data yang akurat dalam mengukur curah hujan [19, 20] sekalipun data tersebut telah direkonstruksi dengan metode pengisian data kosong.

Perlu dipahami bahwa adanya selisih dari perbandingan dua parameter yang sama adalah hal yang wajar dan tidak bisa dihindari karena berbeda jenis alat, metode pengukuran, dan lokasi yang juga tidak sama persis [21]. Selisih rata-rata (MAE) <1 mm yang diperoleh paska pemulihan data menandakan hasil yang sangat signifikan. Ditambah lagi nilai koefisien korelasi yang mendekati 1 semakin menguatkan hubungan linier data otomatis dan manual. Nilai ini mengungguli beberapa metode pengisian *missing data* yang dilakukan pada kajian-kajian terdahulu [22, 23, 24]. Keunggulan ini diyakini karena nilai-nilai yang hilang pada penelitian ini dibangun dari hasil rekaman datanya sendiri sementara penelitian lain dibangun melalui masukan data dari tempat lain [22, 23] ataupun dari alat pengukuran jenis lain [24].

Curah hujan ARG pada umumnya lebih rendah (*underestimate*) dari alat manualnya sebagai representasi nilai slope yang positif (lihat Tabel 3). Penelitian lain menemukan hal yang sama bahwa alat penakar hujan otomatis cenderung lebih *underestimate* [25]. Alat penakar hujan yang menggunakan *tipping bucket* seperti ARG rentan kehilangan besaran hujan karena bejana *tipping bucket* tidak berjungkit dengan baik terutama saat kejadian hujan lebat [20].

Keberhasilan pemulihan data kosong ini membuktikan bahwa data ARG dengan format akumulasi pada menit berikutnya sangat baik dan cocok karena memungkinkan penanganan sederhana pada data hilang dengan akurasi yang tinggi. Permasalahan jaringan yang bersifat sementara tidak akan menjadi permasalahan besar. Apalagi beberapa kasus jaringan bermasalah sering terjadi pada kondisi hujan lebat. Jika terjadi hujan lebat lalu jaringan eror maka ketika hujan telah berkurang atau reda, dan jaringan kembali membaik, data yang sebelumnya tidak terekam di web akan kembali merekam dengan nilai yang telah terakumulasi. Berbeda dengan tipe ARG yang menampilkan data *real* setiap 10 menit akan cenderung lebih rentan menghasilkan nilai harian yang jauh lebih rendah dari sebenarnya. Curah hujan harian ARG dengan tipe tersebut diperoleh dari akumulasi curah hujan selama 24 jam penuh (dari jam 23.00 UTC s.d 23.00 UTC hari berikutnya). Gangguan 10 menit saja selama kejadian hujan maka akan kehilangan data *real* 10 menit tersebut. Apalagi kehilangan data yang lebih lama lagi maka nilai hariannya benar-benar tidak dapat digunakan.

Tipe ARG yang mengakumulasikan curah hujannya ke menit berikutnya membutuhkan koneksi yang baik pada pukul 22.00 hingga 00.00 UTC (untuk wilayah zona waktu WITA) setiap harinya. Oleh karenanya perlu diperhatikan pemasangan alat penakar hujan otomatis ARG di lokasi yang selama rentang waktu tersebut memiliki jaringan yang baik. Telah diketahui bahwa banyak wilayah di Sulawesi Tenggara yang seringkali mengalami gangguan jaringan di waktu-waktu seragam setiap hari. Misalnya jaringan jelek setiap malam hari dan kembali membaik di pagi hari. Pemilihan lokasi yang pas akan menentukan ketersediaan data curah hujan itu sendiri.

4. Kesimpulan

Mengubah data curah hujan ARG dari periode 10 menit menjadi harian dan memulihkan kembali data curah hujan harian kosong menggunakan bahasa pemrograman Python telah berhasil dilakukan. Hasil penelitian menunjukkan persentase *missing data* yang mulanya 21.4% kini tersisa 1.1% saja. Hasil uji validasi data setelah dipulihkan dibandingkan dengan data dari alat manual di lokasi yang sama menunjukkan nilai eror yang rendah (MAE=0.998 mm, RMSE=2.253 mm) dan korelasi yang sangat tinggi (r=0.966). Nilai-nilai ini lebih baik dibandingkan sebelum dilakukan pengisian data kosong. Keberhasilan ini secara tidak langsung membuktikan beberapa hal. Yang pertama bahwa alat otomatis dapat menghasilkan data yang akurat sehingga layak digunakan untuk menggantikan alat manual. Kedua bahwa pengisian data kosong dapat meningkatkan akurasi dan daya guna data. Ketiga bahwa data ARG dengan format akumulasi pada menit berikutnya sangat baik dan cocok karena memungkinkan penanganan sederhana pada data hilang dengan akurasi yang tinggi. Terakhir bahwa perlu diperhatikan pemasangan alat penakar hujan otomatis ARG di lokasi yang selama rentang waktu 22.00 s.d 00.00 UTC memiliki jaringan yang baik. Dengan semakin tingginya persentase kelengkapan data maka penggunaannya akan menjadi semakin luas sehingga memberikan lebih banyak manfaat terutama untuk kebutuhan analisis, prakiraan, pelayanan data, maupun penelitian.

p-ISSN: 2086-5589
e-ISSN: 2723-2239

Ucapan Terima Kasih

Ucapan terima kasih diberikan kepada Pusdiklat BMKG atas diklat yang diberikan terkait bahasa pemrograman python dan bpk. Marjuki, M.Si selaku *coach* selama penyusunan laporan implementasi diklat tersebut.

Daftar Pustaka

[1] BMKG. (2016). *Peraturan Kepala Badan Meteorologi Klimatologi Geofisika No. 4 Tahun 2016 tentang Pengamatan dan Pengelolaan Data Iklim di Lingkungan BMKG*. Retrieved from https://gawbkt.id/assets/jdih/Perka_nomor_4_2016.PDF.

[2] Muita, R., Kucera, P., Aura, S., Muchemi, D., Gikungu, D., Mwangi, S., ... & Kamau, M. (2021). Towards Increasing Data Availability for Meteorological Services: Inter-Comparison of Meteorological Data from a Synoptic Weather Station and Two Automatic Weather Stations in Kenya. *American Journal of Climate Change*, *10*(3), 300-316.

[3] Salsabila, S. (2020). *Materi Modul Online Data Mining Praproses Data Sesi Online 6*. Retrieved from https://lms-paralel.esaunggul.ac.id.

[4] Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*, 1-17.

[5] Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, *17*(1), 1-10.

[6] Alice, M. (2018). *Imputing missing data with R; MICE package*. Retrieved from https://datascienceplus.com/imputing-missing-data-with-r-mice-package/.

[7] Scheffer, J. (2002). Dealing with missing data. *Research Letters in the information and Mathematical Sciences, 3* (1), 153-160

[8] Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of clinical epidemiology*, *110*, 63-73.

[9] Ocampo-Marulanda, C., Cerón, W. L., Avila-Diaz, A., Canchala, T., Alfonso-Morales, W., Kayano, M. T., & Torres, R. R. (2021). Missing data estimation in extreme rainfall indices for the Metropolitan area of Cali-Colombia: An approach based on artificial neural networks. *Data in Brief*, *39*, 107592.

[10] World Meteorological Organization. (2017). *WMO guidelines on the calculation of climate normals*. Geneva, Switzerland: World Meteorological Organization

[11] Miró, J. J., Caselles, V., & Estrela, M. J. (2017). Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric research*, *197*, 313-330.

[12] Muflihah, Pahlawan, R.Y. (2017). Perbandingan Teknik Interpolasi Berbasis R dalam Estimasi Data Curah Hujan Bulanan yang Hilang di Sulawesi. *Jurnal Meteorologi dan Geofisika, 18*(3), 107-111.

[13] Papailiou, I., Spyropoulos, F., Trichakis, I., & Karatzas, G. P. (2022). Artificial Neural Networks and Multiple Linear Regression for Filling in Missing Daily Rainfall Data. *Water*, *14*(18), 2892.

[14] Van Buuren, S. (2018). *Flexible imputation of missing data second edition.* Vancouver, Canada :CRC press Taylor & Francis Group.

[15] Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, *4*(1), 1 - 8

[16] Koop, D. (2021). *Advanced Data Management (CSCI 490/680).* Retrieved from https://faculty.cs.niu.edu/~dakoop/

[17] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82.

[18] Kurniawan, D. (2008). *Regresi Linier*. Retrieved from https://ineddeni.files.wordpress.com/2008/07/regresi_linier.pdf.

[19] Risnayah, S. (2022). Uji Keakuratan Data Suhu Udara, Kelembaban Udara, Tekanan Udara, dan Curah Hujan dari Alat Automatic Weather Station terhadap Pengukuran Manualnya. *Megasains*, *13*(2), 18-25

[20] Wicaksana, H. S., & Putra, M. (2021). *Evaluasi Kinerja Automatic Weather Station Berdasarkan Pengamatan Paralel di Stasiun Meteorologi Kemayoran*. Prosiding Seminar Nasional Teknik Elektro (pp.59-64). Politeknik Negeri Jakarta, Indonesia

[21] Xiaohui, W. Y. L. X. J. (2006). *Differences between automatic and manual meteorological observation.* TECO-2006-WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation. Geneva, Switzerland

[22] Prawaka, F., Zakaria, A., & Tugiono, S. (2016). Analisis Data Curah Hujan yang Hilang Dengan Menggunakan Metode Normal Ratio, Inversed Square Distance, dan Rata-Rata Aljabar (Studi Kasus Curah Hujan Beberapa Stasiun Hujan Daerah Bandar Lampung). *Jurnal Rekayasa Sipil dan Desain* 4(3), 397-406.

[23] Kurniawan, R. D. (2017). *Mengisi Data Hujan yang Hilang dengan Metode Autoregressive dan Metode Reciprocal dengan Pengujian Debit Kala Ulang (Studi Kasus di DAS Bakalan).* Surakarta: Universitas Sebelas Maret.

[24] Duarte, L. V., Formiga, K. T. M., & Costa, V. A. F. (2022). Comparison of Methods for Filling Daily and Monthly Rainfall Missing Data: Statistical Models or Imputation of Satellite Retrievals?. *Water*, *14*(19), 3144.

[25] Valík, A., Brázdil, R., Zahradníček, P., Tolasz, R., & Fiala, R. (2021). Precipitation measurements by manual and automatic rain gauges and their influence on homogeneity of long-term precipitation series. *International Journal of Climatology*, *41*, E2537-E2552.

p-ISSN: 2086-5589
e-ISSN: 2723-2239