

Penerapan K-Means Clustering untuk Identifikasi Pola Iklim Stasiun Cuaca Indonesia Menggunakan Data NOAA Global Surface Summary of the Day (GSOD)

Andly Sofian Hasugian^{1*}, Jatmiko Althaf Aziz², Nayla Anjani Nasution³

^{1,2,3}Matematika dan Ilmu Pengetahuan Alam, Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia

Email: ^{1*}dlyhasugian@gmail.com, ²althafaziz02@gmail.com, ³naylanjaninst26@gmail.com

(*Email Corresponding Author: dlyhasugian@gmail.com)

Received: 18 Maret 2026 | Revision: 23 Maret 2026 | Accepted: 9 Juni 2026

Abstrak

Perubahan iklim yang terjadi secara global menuntut adanya analisis data cuaca yang lebih sistematis dan terstruktur, khususnya untuk wilayah Indonesia yang memiliki keragaman iklim tinggi akibat pengaruh geografis kepulauan. Penelitian ini menerapkan algoritma K-Means Clustering untuk mengidentifikasi pola iklim stasiun cuaca di Indonesia menggunakan dataset NOAA Global Surface Summary of the Day (GSOD) yang diakses secara langsung melalui Application Programming Interface (API) resmi NOAA NCEI. Sebanyak 87 stasiun cuaca aktif di seluruh Indonesia digunakan sebagai objek penelitian dengan periode pengamatan tahun 2023, menghasilkan 50.232 records data mentah. Variabel yang digunakan mencakup suhu maksimum, suhu minimum, suhu rata-rata harian, dan curah hujan. Normalisasi data dilakukan menggunakan metode StandardScaler sebelum proses clustering. Penentuan jumlah cluster optimal menggunakan Elbow Method menghasilkan nilai $K=4$, ditandai dengan penurunan inertia (WCSS) yang signifikan dari 72.33 ($K=4$) dibanding 132.02 ($K=3$). Hasil clustering mengelompokkan 87 stasiun ke dalam empat cluster: Cluster 0-Tropis Basah Sedang (38 stasiun, 43.7%), Cluster 1-Dataran Tinggi Sejuk (4 stasiun, 4.6%), Cluster 2-Tropis Panas Basah (43 stasiun, 49.4%), dan Cluster 3-Tropis Curah Hujan Ekstrem (2 stasiun, 2.3%). Hasil penelitian divisualisasikan dalam peta interaktif berbasis Folium yang menampilkan distribusi spasial cluster secara geografis. Pendekatan berbasis API ini membuktikan efisiensi akses big data skala petabyte tanpa proses unduhan massal.

Kata Kunci: K-Means Clustering, NOAA GSOD, Pola Iklim, Stasiun Cuaca Indonesia, Elbow Method, Big Data API

Abstract

Global climate change demands more systematic and structured weather data analysis, especially for Indonesia which has high climate diversity due to its archipelagic geographical influence. This study applies the K-Means Clustering algorithm to identify climate patterns of weather stations in Indonesia using the NOAA Global Surface Summary of the Day (GSOD) dataset accessed directly through the official NOAA NCEI Application Programming Interface (API). A total of 87 active weather stations across Indonesia were used as research objects with an observation period of 2023, yielding 50,232 raw data records. Variables used include maximum temperature, minimum temperature, daily average temperature, and precipitation. Data normalization was performed using the StandardScaler method prior to clustering. Determination of the optimal number of clusters using the Elbow Method yielded $K=4$, marked by a significant decrease in inertia (WCSS) from 72.33 ($K=4$) compared to 132.02 ($K=3$). Clustering results grouped 87 stations into four clusters: Cluster 0-Moderate Tropical Wet (38 stations, 43.7%), Cluster 1-Highland Cool (4 stations, 4.6%), Cluster 2-Hot Tropical Wet (43 stations, 49.4%), and Cluster 3-Tropical Extreme Rainfall (2 stations, 2.3%). Results are visualized in a Folium-based interactive map displaying the geographic spatial distribution of clusters. This API-based approach demonstrates the efficiency of accessing petabyte-scale big data without mass downloads.

Keywords: K-Means Clustering, NOAA GSOD, Climate Pattern, Indonesian Weather Stations, Elbow Method, Big Data API

1. PENDAHULUAN

Perubahan iklim global yang terjadi secara masif telah mendorong kebutuhan analisis data cuaca dalam skala besar secara lebih efisien dan komprehensif. Indonesia sebagai negara kepulauan terbesar di dunia dengan letak geografis di kawasan tropis memiliki karakteristik iklim yang sangat beragam, dipengaruhi oleh faktor topografi, jarak dari laut, pola angin musiman, serta fenomena El Niño dan La Niña [1]. Keberagaman iklim ini menjadi tantangan tersendiri dalam memahami distribusi spasial pola cuaca di seluruh wilayah nusantara secara sistematis dan terukur.

Data cuaca dalam jumlah besar (big data) kini tersedia secara publik melalui berbagai platform cloud internasional. Salah satu sumber data cuaca paling komprehensif di dunia adalah NOAA (National Oceanic and Atmospheric Administration), lembaga pemerintah Amerika Serikat yang menyediakan dataset Global Surface Summary of the Day (GSOD) yang mencakup lebih dari 9.000 stasiun cuaca di seluruh dunia sejak tahun 1929 hingga saat ini dengan total ukuran data mencapai 24 Petabyte (PB) [2]. Keunggulan dataset ini dibanding sumber data lain adalah cakupan globalnya yang luas, konsistensi format data, dan aksesibilitasnya melalui Application Programming Interface (API) tanpa perlu mengunduh keseluruhan dataset.

Berbagai penelitian telah dilakukan terkait analisis clustering data cuaca di Indonesia. Manalu et al. [3] menerapkan K-Means Clustering untuk menganalisis pola cuaca di Provinsi Sumatera Utara dan berhasil mengidentifikasi kelompok-kelompok wilayah berdasarkan karakteristik cuaca yang berbeda. Penelitian lain oleh Lewenusa dan Handhayani [4] menggunakan kombinasi K-Means dan Fuzzy C-Means (FCM) pada data meteorologi BMKG dari 28 stasiun di Indonesia Timur untuk mendeteksi tanda-tanda perubahan iklim, dan menemukan adanya tren kenaikan temperatur yang signifikan. Rahman et al. [5] menerapkan K-Means Clustering pada data iklim Provinsi Jambi periode 2020-2024 menggunakan Elbow Method untuk menentukan K optimal dan berhasil mengidentifikasi tiga rezim iklim berbeda. Chusyairi [6] melakukan clustering data cuaca ekstrem Indonesia menggunakan K-Means dan entropi, menghasilkan 3 kelas tingkat cuaca dengan menghindari bias terhadap presisi non-optimal. Fadil et al. [7] mengkombinasikan K-Means dengan Random Forest pada data cuaca Cilacap BMKG dari tahun 1975-2023, membuktikan efektivitas clustering sebagai preprocessing untuk klasifikasi cuaca.

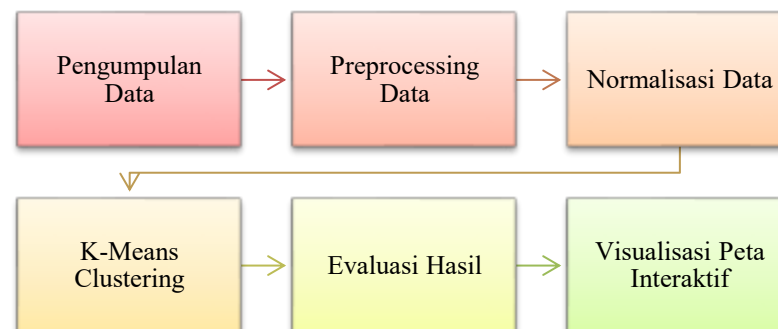
Berdasarkan kajian penelitian terdahulu, terdapat beberapa gap yang belum terpenuhi: (1) belum ada penelitian yang mengakses data NOAA GSOD secara langsung melalui API untuk analisis clustering stasiun cuaca Indonesia; (2) sebagian besar penelitian terbatas pada satu wilayah provinsi atau pulau tertentu, belum mencakup seluruh Indonesia; (3) belum ada penelitian yang mengintegrasikan visualisasi peta interaktif Folium sebagai output analisis clustering cuaca berbasis API. Penelitian ini hadir untuk mengisi gap tersebut dengan memanfaatkan paradigma akses big data berbasis API yang lebih efisien.

Tujuan penelitian ini adalah: (1) mengakses dan mengolah data NOAA GSOD secara real-time melalui API NCEI tanpa proses unduhan massal; (2) menerapkan algoritma K-Means Clustering pada 87 stasiun cuaca Indonesia; (3) menentukan jumlah cluster optimal menggunakan Elbow Method; dan (4) memvisualisasikan hasil clustering dalam bentuk peta interaktif dan tabel ringkasan yang dapat diinterpretasikan secara geografis. Penelitian ini diharapkan dapat berkontribusi pada pemahaman pola iklim Indonesia dan menjadi referensi metode akses big data cuaca berbasis API untuk penelitian selanjutnya.

2. METODOLOGI PENELITIAN

2.1 Alur Penelitian

Penelitian ini dilaksanakan melalui enam tahapan sistematis seperti ditunjukkan pada Gambar 1. Setiap tahapan dirancang untuk memastikan kualitas data dan hasil clustering yang valid secara klimatologis.



Gambar 1. Diagram Alur Tahapan Penelitian

Tahapan penelitian dimulai dari pengumpulan data melalui API NOAA, dilanjutkan preprocessing, normalisasi, penentuan K optimal dengan Elbow Method, penerapan K-Means Clustering, evaluasi hasil, hingga visualisasi dalam peta interaktif.

2.2 Pengumpulan Data melalui API NOAA

Data diperoleh dari NOAA NCEI melalui CDO Web Services API v2 dengan dataset identifier GHCND. Proses pengambilan data menggunakan parameter locationid=FIPS:ID untuk stasiun Indonesia pada periode 1 Januari 2023 hingga 31 Desember 2023. Total data mentah yang dikumpulkan berjumlah 50.232 records dari 91 stasiun terdaftar, dengan 87 stasiun memiliki data lengkap.

Tabel 1. Atribut Dataset NOAA GSOD yang Digunakan

No	Kode API	Keterangan	Satuan	Sumber
1	TMAX	Suhu Maksimum Harian	°C	GHCND
2	TMIN	Suhu Minimum Harian	°C	GHCND
3	TOBS	Suhu Rata-rata Harian	°C	Turunan
4	PRCP	Curah Hujan Harian	mm	GHCND

```

1  import requests
2  import pandas as pd
3  import time
4
5  TOKEN = 'token_noaa_nya'
6  BASE_URL = 'https://www.ncdc.noaa.gov/cdo-web/api/v2'
7  HEADERS = {'token': TOKEN}
8
9  # mengambil daftar stasiun cuaca di Indonesia
10 params = {
11     'datasetid': 'GHCND',
12     'locationid': 'FIPS:ID', # ini adalah kode lokasi Indonesia
13     'limit': 100
14 }
15 response = requests.get(f'{BASE_URL}/stations', headers=HEADERS, params=params)
16 df_stations = pd.DataFrame(response.json()['results'])

```

Gambar 2. Implementasi Kode Akses API NOAA NCEI

2.3 Preprocessing Data

Data mentah berbentuk format panjang (long format) diproses melalui: (1) pivot tabel mengubah format menjadi wide format; (2) agregasi nilai rata-rata tahunan per stasiun; (3) penanganan missing value dengan median kolom; dan (4) normalisasi data menggunakan StandardScaler (Z-Score) sebelum clustering.

2.4 Normalisasi Data

Normalisasi data diperlukan sebelum proses clustering karena atribut dengan skala yang berbeda dapat mendominasi perhitungan jarak Euclidean. StandardScaler melakukan standarisasi fitur dengan menghapus rata-rata dan menskalakan ke varians unit, menggunakan rumus Z-score:

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

Dimana x adalah nilai asli, μ adalah rata-rata kolom, dan σ adalah standar deviasi kolom. Setelah normalisasi, setiap fitur memiliki rata-rata 0 dan standar deviasi 1, sehingga tidak ada fitur yang mendominasi perhitungan jarak [9].

2.5 Penerapan K-Means Clustering

K-Means Clustering mengelompokkan data ke dalam K cluster berdasarkan kemiripan jarak Euclidean. Jarak dihitung menggunakan:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2} \quad (2)$$

Fungsi objektif yang diminimalkan adalah Within-Cluster Sum of Squares (WCSS):

$$J = \sum_k \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (3)$$

Centroid diperbarui setiap iterasi menggunakan:

$$\mu_k = \frac{1}{|C_k| \sum_{x_i \in C_k} x_i} \quad (3)$$

Algoritma dijalankan menggunakan scikit-learn dengan parameter $n_clusters=4$, $random_state=42$, dan $n_init='auto'$ (10 kali inisialisasi, dipilih hasil inertia terkecil). Inisialisasi centroid awal menggunakan metode K-Means++ yang memilih centroid berdasarkan probabilitas proporsional terhadap kuadrat jarak ke centroid terdekat, sehingga tersebar merata dan mempercepat konvergensi.

2.6 Implementasi Perhitungan Manual dengan Data Nyata

Bagian ini menunjukkan implementasi seluruh rumus K-Means Clustering secara step-by-step menggunakan data stasiun cuaca Indonesia yang sebenarnya. Tujuannya adalah membuktikan kesesuaian hasil komputasi Python dengan perhitungan matematis manual.

2.6.1 Langkah 1 — Normalisasi Data (Z-Score)

Sebelum clustering, seluruh fitur dinormalisasi menggunakan Z-score agar tidak ada fitur yang mendominasi perhitungan. Sebelum clustering, seluruh fitur dinormalisasi menggunakan Z-score agar tidak ada fitur yang mendominasi perhitungan jarak. Parameter normalisasi dari 87 stasiun: μ Suhu Max = 31.9339°C , σ Suhu Max = 1.5133 ; μ Suhu Min = 23.5767°C , σ Suhu Min = 1.6454 ; μ Suhu Rata² = 27.7556°C , σ Suhu Rata² = 1.4481 ; μ Curah Hujan = 13.2663 mm, σ Curah Hujan = 13.0119 .

Contoh perhitungan normalisasi untuk Stasiun Jakarta Observatory (Suhu Max= 32.70 , Suhu Min= 25.94 , Suhu Rata²= 29.32 , Hujan= 0.50): $z_suhu_max = (32.70 - 31.9339) / 1.5133 = 0.5062$; $z_suhu_min = (25.94 - 23.5767) / 1.6454 = 1.4363$; $z_suhu_rata = (29.32 - 27.7556) / 1.4481 = 1.0803$; $z_hujan = (0.5 - 13.2663) / 13.0119 = -0.9811$.

Sehingga vektor fitur Jakarta Observatory setelah normalisasi adalah: $x = [0.5062, 1.4363, 1.0803, -0.9811]$. Implementasi Python:

```
from sklearn.preprocessing import StandardScaler

fitur = ['Suhu Max (°C)', 'Suhu Min (°C)', 'Suhu Rata-Rata (°C)', 'Curah Hujan (mm)']

scaler = StandardScaler()
x_scaled = scaler.fit_transform(df_stations[fitur])

# Hasil: mean dan std yang digunakan
# mean = [31.9339, 23.5767, 27.7556, 13.2663]
# std = [1.5133, 1.6454, 1.4481, 13.0119]
```

Gambar 3. Implementasi Normalisasi Z-Score (StandardScaler)

2.6.2 Langkah 2 — Centroid Setelah Konvergensi

Setelah algoritma K-Means konvergen, posisi centroid akhir masing-masing cluster dalam skala data asli: μ_0 (Cluster 0 - Tropis Basah Sedang) = $[31.42, 24.00, 27.71, 10.94]$; μ_1 (Cluster 1 - Dataran Tinggi Sejuk) = $[26.51, 17.10, 21.80, 7.75]$; μ_2 (Cluster 2 - Tropis Panas Basah) = $[32.89, 23.82, 28.36, 12.52]$; μ_3 (Cluster 3 - Curah Hujan Ekstrem) = $[31.91, 23.34, 27.62, 84.58]$. Format vektor: [Suhu Max ($^{\circ}\text{C}$), Suhu Min ($^{\circ}\text{C}$), Suhu Rata² ($^{\circ}\text{C}$), Curah Hujan (mm)].

Centroid Cluster 1 memiliki suhu jauh lebih rendah (21.80°C) dibanding cluster lain, mencerminkan karakteristik dataran tinggi. Cluster 3 memiliki curah hujan ekstrem (84.58 mm/hari), hampir 7 kali lipat rata-rata nasional.

2.6.3 Langkah 3 — Perhitungan Jarak Euclidean

Penugasan setiap stasiun ke cluster dilakukan berdasarkan jarak Euclidean minimum ke centroid terdekat. Contoh perhitungan untuk Stasiun Jakarta Observatory $x = [0.5062, 1.4363, 1.0803, -0.9811]$:

$$d(\text{Jakarta}, C_0) = \sqrt{(0.5062 - (-0.3274))^2 + (1.4363 - 0.2587)^2 + (1.0803 - (-0.0301))^2 + (-0.9811 - (-0.1786))^2} = \sqrt{0.6949 + 1.3863 + 1.2330 + 0.6440} = \sqrt{3.9582} = 1.9959$$

$$d(\text{Jakarta}, C_1) = 8.5364 \quad | \quad d(\text{Jakarta}, C_2) = 1.7259 \quad | \quad d(\text{Jakarta}, C_3) = 6.7741$$

Karena $d(\text{Jakarta}, C_2) = 1.7259$ adalah nilai terkecil, Jakarta Observatory masuk Cluster 2 (Tropis Panas Basah). Contoh lain untuk Stasiun Enarotali $x = [-4.9667, -5.1656, -5.5109, -0.2000]$:

$$d(\text{Enarotali}, C_0) = 9.0234 \quad | \quad d(\text{Enarotali}, C_1) = 2.3605 \quad | \quad d(\text{Enarotali}, C_2) = 9.7623 \quad | \quad d(\text{Enarotali}, C_3) = 10.5809$$

Karena $d(\text{Enarotali}, C_1) = 2.3605$ adalah nilai terkecil, Enarotali masuk Cluster 1 (Dataran Tinggi Sejuk).

2.6.4 Langkah 4 — Perhitungan WCSS / Inertia

Kualitas clustering diukur menggunakan Within-Cluster Sum of Squares (WCSS). Nilai total WCSS diperoleh dengan menjumlahkan WCSS setiap cluster:

$$J(K=4) = WCSS_0 + WCSS_1 + WCSS_2 + WCSS_3 = 26.1587 + 14.4021 + 30.3047 + 1.4617 = 72.3272$$

Rincian: WCSS₀ = 26.1587 (38 stasiun, rata-rata 0.6884 per stasiun); WCSS₁ = 14.4021 (4 stasiun, rata-rata 3.6005 per stasiun); WCSS₂ = 30.3047 (43 stasiun, rata-rata 0.7048 per stasiun); WCSS₃ = 1.4617 (2 stasiun, rata-rata 0.7309 per stasiun).

Cluster 1 memiliki WCSS per stasiun tertinggi (3.6005) karena terdiri dari 4 stasiun dataran tinggi yang tersebar di lokasi berbeda (Kerinci 782m, Bogor, Wamena 1660m, Enarotali 1770m). Cluster 0 dan Cluster 2 memiliki WCSS per stasiun paling rendah (~0.70), menunjukkan stasiun tropis dataran rendah sangat homogen.

3. HASIL DAN PEMBAHASAN

3.1 Statistik Deskriptif Dataset

Tabel 2 menyajikan statistik deskriptif dari 87 stasiun cuaca Indonesia yang berhasil dikumpulkan melalui API NOAA untuk periode tahun 2023. Rata-rata suhu maksimum seluruh stasiun adalah 31.93°C dengan standar deviasi 1.52°C, menunjukkan variasi suhu yang relatif rendah — mencerminkan karakteristik iklim tropis Indonesia yang cenderung homogen. Namun, variabel curah hujan menunjukkan variasi yang jauh lebih tinggi dengan standar deviasi 13.09 mm, rentang antara 0.40 mm hingga 93.70 mm per hari, mengindikasikan keragaman intensitas hujan yang signifikan antar wilayah.

Tabel 2. Statistik Deskriptif Dataset NOAA GSOD Stasiun Cuaca Indonesia 2023

Statistik	Suhu Max (°C)	Suhu Min (°C)	Suhu Rata ² (°C)	Curah Hujan (mm)
Jumlah (n)	87	87	87	87
Rata-rata	31.93	23.58	27.76	13.27
Std. Deviasi	1.52	1.65	1.46	13.09
Minimum	24.41	15.05	19.73	0.40
Q1 (25%)	31.47	23.29	27.64	6.74
Median	32.16	23.80	28.01	10.64
Q3 (75%)	32.76	24.26	28.37	14.62
Maksimum	34.18	25.94	29.32	93.70

Stasiun dengan suhu tertinggi adalah Jakarta Observatory (29.32°C rata-rata tahunan), diikuti Semarang (29.31°C) dan Surabaya Perak (29.26°C) — ketiganya merupakan kota-kota besar di dataran rendah Pulau Jawa. Sebaliknya, stasiun terdingin adalah Enarotali di Papua (19.73°C) dengan elevasi 1.770 meter, diikuti Wamena (20.96°C, elevasi 1.660 m) dan Bogor Citeko (22.46°C). Pola ini mengkonfirmasi pengaruh elevasi yang kuat terhadap suhu udara di Indonesia.

3.2 Hasil Elbow Method

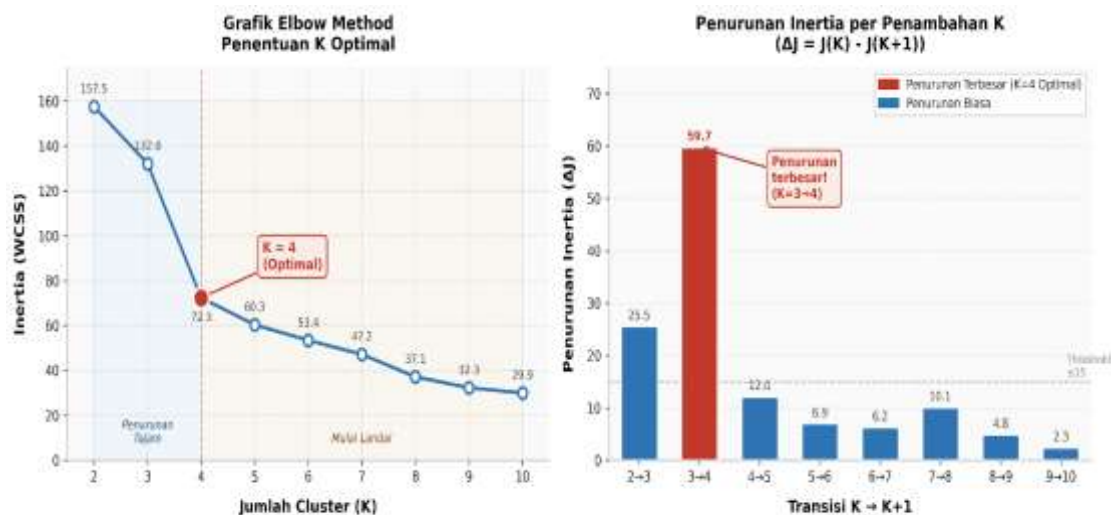
Tabel 3 menampilkan nilai inertia (WCSS) untuk setiap nilai K dari 2 hingga 10. Penurunan inertia yang paling signifikan terjadi dari K=3 ke K=4, yaitu sebesar 59.69 poin (dari 132.02 menjadi 72.33). Pada K=4 ke K=5, penurunan inertia hanya sebesar 12.04 poin, jauh lebih kecil dibanding iterasi sebelumnya. Fenomena ini menandai titik 'siku' yang jelas pada K=4, sehingga K=4 dipilih sebagai jumlah cluster optimal.

Tabel 3. Nilai Inertia Elbow Method untuk K=2 hingga K=10

K	Inertia (WCSS)	Keterangan
2	157.5164	
3	132.0165	
4	72.3272	← Optimal (Elbow)

K	Inertia (WCSS)	Keterangan
5	60.2910	
6	53.3632	
7	47.1571	
8	37.1038	
9	32.2603	
10	29.9304	

Hasil ini sejalan dengan penelitian Rahman et al. [5] yang menemukan $K=3$ sebagai optimal untuk data iklim Jambi, dan Manalu et al. [3] yang menemukan K optimal untuk data cuaca Sumatera Utara. Perbedaan nilai K optimal yang wajar mengingat cakupan geografis penelitian ini lebih luas (seluruh Indonesia) dibanding penelitian-penelitian tersebut yang berfokus pada satu provinsi.



Gambar 3. Grafik Elbow Method — Penentuan Jumlah Cluster Optimal ($K=4$) pada Data Stasiun Cuaca Indonesia

Gambar 4. Grafik Elbow Method — Penentuan Jumlah Cluster Optimal ($K=4$) pada Data Stasiun Cuaca Indonesia 2023

Interpretasi grafik menunjukkan bahwa pada $K=4$, algoritma K-Means telah berhasil menemukan struktur natural dalam data — yaitu empat pola iklim yang memang berbeda secara klimatologis di Indonesia: iklim tropis dataran rendah (Cluster 0 dan 2), iklim dataran tinggi (Cluster 1), dan iklim dengan curah hujan ekstrem (Cluster 3). Penambahan cluster melebihi $K=4$ hanya akan memecah kelompok-kelompok yang sudah terbentuk secara alami ini menjadi sub-kelompok yang kurang bermakna secara meteorologis.

3.3 Hasil K-Means Clustering

Dengan $K=4$, algoritma K-Means berhasil mengelompokkan 87 stasiun cuaca Indonesia ke dalam empat cluster dengan karakteristik yang berbeda dan bermakna secara meteorologis. Tabel 4 menyajikan ringkasan karakteristik setiap cluster.

Tabel 4. Ringkasan Karakteristik Hasil K-Means Clustering ($K=4$)

Cluster	Nama Cluster	Jml Stasiun	Suhu Max (°C)	Suhu Min (°C)	Suhu Rata ² (°C)	Hujan (mm)
0	Tropis Basah Sedang	38 (43.7%)	31.42	24.00	27.71	10.94
1	Dataran Tinggi Sejuk	4 (4.6%)	26.51	17.10	21.80	7.75

Cluster	Nama Cluster	Jml Stasiun	Suhu Max (°C)	Suhu Min (°C)	Suhu Rata ² (°C)	Hujan (mm)
2	Tropis Panas Basah	43 (49.4%)	32.89	23.82	28.36	12.52
3	Tropis Curah Hujan Ekstrem	2 (2.3%)	31.91	23.34	27.62	84.58

Cluster 2 (Tropis Panas Basah) merupakan cluster terbesar dengan 43 stasiun (49.4%), dicirikan oleh suhu rata-rata tertinggi (28.36°C) dan curah hujan tinggi (12.52 mm/hari). Stasiun dalam cluster ini umumnya berada di dataran rendah pesisir Jawa, Sulawesi, dan Papua, mencerminkan karakteristik iklim tropis panas basah yang mendominasi Indonesia bagian tengah dan timur.

Cluster 0 (Tropis Basah Sedang) terdiri dari 38 stasiun (43.7%) dengan suhu rata-rata 27.71°C dan curah hujan 10.94 mm/hari. Cluster ini mencakup sebagian besar stasiun di Sumatera, Kalimantan, dan Indonesia bagian timur, dengan elevasi rata-rata rendah (25.63 m). Perbedaan antara Cluster 0 dan Cluster 2 terletak pada intensitas suhu dan curah hujannya — Cluster 2 lebih panas dan lebih basah secara konsisten.

Cluster 1 (Dataran Tinggi Sejuk) adalah cluster terkecil dengan hanya 4 stasiun (4.6%), namun memiliki karakteristik yang paling distinktif. Elevasi rata-rata mencapai 1.404 meter, dengan suhu rata-rata jauh lebih rendah (21.80°C) dan suhu minimum mencapai 15.05°C di Enarotali. Keempat stasiun ini adalah Kerinci (782 m), Bogor Citeko, Wamena (1.660 m), dan Enarotali (1.770 m) — semuanya merupakan stasiun di dataran tinggi. Hasil ini mengkonfirmasi pengaruh dominan elevasi terhadap pembentukan zona iklim sejuk di Indonesia.

Cluster 3 (Tropis Curah Hujan Ekstrem) hanya terdiri dari 2 stasiun dengan karakteristik unik: curah hujan rata-rata 84.58 mm/hari — hampir 7 kali lipat dibanding cluster lain. Dua stasiun ini adalah Naha (93.70 mm/hari) dan Manado Sam Ratulangi (75.45 mm/hari), keduanya berlokasi di Sulawesi Utara. Fenomena curah hujan ekstrem ini disebabkan oleh posisi geografis Sulawesi Utara yang menghadap langsung ke Samudra Pasifik, pengaruh angin muson, dan topografi pegunungan yang mendorong curah hujan orografis.

Tabel 5. Sampel Stasiun per Cluster Hasil K-Means

Nama Stasiun	Suhu Rata ² (°C)	Hujan (mm)	Cluster
Cluster 0 — Tropis Basah Sedang (38 stasiun)			
Medan Belawan	28.46	10.64	0
Pangkalpinang	28.10	14.70	0
Ambon Pattimura	27.22	12.16	0
Cluster 1 — Dataran Tinggi Sejuk (4 stasiun)			
Kerinci Depati Parbo	24.06	6.16	1
Wamena	20.96	7.40	1
Enarotali	19.73	10.64	1
Cluster 2 — Tropis Panas Basah (43 stasiun)			
Jakarta Observatory	29.32	0.50	2
Semarang	29.31	33.55	2
Surabaya Perak	29.26	10.64	2
Cluster 3 — Tropis Curah Hujan Ekstrem (2 stasiun)			
Naha	27.63	93.70	3

3.4 Visualisasi Peta Interaktif

Hasil clustering divisualisasikan menggunakan library Folium yang menghasilkan peta interaktif berbasis OpenStreetMap. Setiap stasiun direpresentasikan sebagai marker berwarna pada peta, dengan warna merah (Cluster 0), biru (Cluster 1), hijau (Cluster 2), dan oranye (Cluster 3).

Peta interaktif yang dihasilkan memungkinkan pengguna melakukan zoom dan mengklik setiap titik stasiun untuk melihat informasi detail. Distribusi spasial cluster pada peta menunjukkan pola geografis yang jelas: Cluster 2 (hijau) mendominasi wilayah pesisir utara Jawa dan dataran rendah Sulawesi; Cluster 0 (merah) tersebar di Sumatera bagian timur dan Kalimantan; Cluster 1 (biru) terisolasi di titik-titik dataran tinggi; dan Cluster 3 (oranye) terkonsentrasi di Sulawesi Utara.



Gambar 5. Peta Distribusi Spasial Hasil K-Means Clustering (K=4) Stasiun Cuaca Indonesia Tahun 2023

4. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma K-Means Clustering untuk mengidentifikasi pola iklim 87 stasiun cuaca Indonesia menggunakan dataset NOAA Global Surface Summary of the Day (GSOD) yang diakses melalui API NCEI secara langsung, tanpa proses unduhan massal dari dataset berukuran 24 Petabyte. Pendekatan berbasis API ini membuktikan efisiensi dan kepraktisan analisis big data cuaca dalam skala nasional. Penerapan Elbow Method menghasilkan K=4 sebagai jumlah cluster optimal, ditandai dengan penurunan inertia (WCSS) signifikan dari 132.02 (K=3) menjadi 72.33 (K=4). Hasil K-Means Clustering mengelompokkan 87 stasiun ke dalam: (1) Cluster 0-Tropis Basah Sedang (38 stasiun, 43.7%, suhu rata² 27.71°C, hujan 10.94 mm); (2) Cluster 1-Dataran Tinggi Sejuk (4 stasiun, 4.6%, suhu rata² 21.80°C, elevasi rata² 1.404 m); (3) Cluster 2-Tropis Panas Basah (43 stasiun, 49.4%, suhu rata² 28.36°C, hujan 12.52 mm); dan (4) Cluster 3-Tropis Curah Hujan Ekstrem (2 stasiun, 2.3%, hujan ekstrem 84.58 mm/hari di Sulawesi Utara). Temuan kunci penelitian ini adalah: dominasi cluster tropis (>93% stasiun), pengaruh kuat elevasi dalam pembentukan Cluster 1 (Dataran Tinggi Sejuk), dan fenomena curah hujan ekstrem Sulawesi Utara yang membentuk cluster tersendiri. Hasil ini berkontribusi pada pemahaman distribusi spasial zona iklim Indonesia yang dapat dimanfaatkan untuk perencanaan pertanian berbasis zona iklim, mitigasi bencana hidrometeorologi, dan kebijakan pengelolaan sumber daya air. Penelitian selanjutnya disarankan untuk: (1) menggunakan rentang waktu lebih panjang (multi-tahun) guna mendeteksi tren perubahan iklim; (2) menambahkan variabel kelembaban dan kecepatan angin; (3) membandingkan hasil K-Means dengan metode clustering lain seperti DBSCAN atau K-Medoids; dan (4) mengintegrasikan data BMKG untuk validasi dan perbandingan hasil clustering.

REFERENCES

- [1] A. Kurniadi, E. Weller, Y. H. Kim, dan S. K. Min, "Evaluation of Coupled Model Intercomparison Project Phase 6 model-simulated extreme precipitation over Indonesia," *International Journal of Climatology*, vol. 43, no. 1, pp. 174-196, Jan. 2023. DOI: 10.1002/joc.7744
- [2] NOAA National Centers for Environmental Information, "Global Surface Summary of the Day (GSOD)," NOAA NCEI, 2023. [Online]. Available: <https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516>
- [3] E. Manalu, E. Surbakti, dan S. P. Sipayung, "Analisis Pola Cuaca di Provinsi Sumatera Utara Menggunakan Metode Clustering K-Means," *Jurnal Ilmu Komputer dan Informatika*, vol. 2, no. 3, pp. 81-89, Jan. 2026.
- [4] I. Lewenusa dan T. Handhayani, "Clustering Data Meteorologi Kota-Kota di Indonesia Timur Menggunakan Metode K-Means dan Fuzzy C-Means," *INTI Nusa Mandiri*, vol. 18, no. 2, pp. 100-106, Feb. 2024. DOI: 10.33480/inti.v18i2.5039
- [5] F. D. Rahman, "Penerapan Algoritma K-Means untuk Klasterisasi Pola Iklim di Provinsi Jambi Menggunakan Elbow Method," *DJTechno*, vol. 6, no. 3, Des. 2025. DOI: 10.46576/djtechno
- [6] A. Chusyairi, "Clustering Data Cuaca Ekstrem Indonesia dengan K-Means dan Entropi," *Journal of Informatics and Communication Technology (JICT)*, 2023. DOI: 10.ejournal.akademitelkom

- [7] F. D. Rahman, M. I. Z. Mulki, dan A. Taryana, "Clustering dan Klasifikasi Data Cuaca Cilacap dengan Menggunakan Metode K-Means dan Random Forest," *Jurnal SINTA: Sistem Informasi dan Teknologi Komputasi*, 2024.
- [8] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281-297, 1967.
- [9] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011. DOI: 10.5555/1953048.2078195
- [10] A. K. Jain, "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010. DOI: 10.1016/j.patrec.2009.09.011
- [11] K. P. Sinaga dan M. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716-80727, 2020. DOI: 10.1109/ACCESS.2020.2988796
- [12] E. Umargono, J. E. Suseno, dan V. Gunawan, "K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula," *Advances in Social Science, Educational and Humanities Research*, vol. 474, pp. 121-129, 2020.
- [13] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, dan B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conference Series*, vol. 336, 2018. DOI: 10.1088/1757-899X/336/1/012017
- [14] M. Fathi, M. Haghi Kashani, S. M. Jameii, dan E. Mahdipour, "Big Data Analytics in Weather Forecasting: A Systematic Review," *Archives of Computational Methods in Engineering*, 2021. DOI: 10.1007/s11831-021-09616-4
- [15] W. McKinney, "Data Structures for Statistical Computing in Python," *Proceedings of 9th Python in Science Conference*, vol. 445, pp. 51-56, 2010. DOI: 10.25080/Majora-92bf1922-00a