

Imputasi Curah Hujan ERA5 Menggunakan *Random Forest* dan *XGBoost* di Maluku Utara

DOI: <http://dx.doi.org/10.35889/progresif.v22i3.3734>

Creative Commons License 4.0 (CC BY –NC)



Balya Badar Syah¹, Brina Miftahurrohmah^{2*}

Sistem Informasi, Universitas Internasional Semen Indonesia, Gresik, Indonesia

*e-mail Corresponding Author: brina.miftahurrohmah@uisi.ac.id

Abstract

Climate change increases hydrometeorological disaster risks in North Maluku, yet missing BMKG rainfall data and ERA5's spatial bias hinder precise risk analysis. This study applied point-based Statistical Downscaling using Random Forest (RF) and Extreme Gradient Boosting (XGBoost), with ERA5 precipitation as the single predictor, to impute missing rainfall at four stations. XGBoost achieved lower MSE and RMSE at three stations (Sultan Babullah: MSE 1017.29, RMSE 31.89, MAE 12.31, R^2 0.0351; Gamar Malamo: MSE 513.96, RMSE 22.67, MAE 9.91, R^2 -0.0141; Oesman Sadik: MSE 163.59, RMSE 12.79, MAE 6.27, R^2 0.0369), with NRMSE-based accuracy of 83.89 – 87.56%, though RF retained slightly lower MAE. RF was selected at Emalamo (MSE 163.75, RMSE 12.79, MAE 5.62, R^2 -0.0517, accuracy 85.78%). The models imputed 298 missing days, though the univariate predictor underestimated rainfall above 40 mm. The resulting continuous dataset offers a scientific basis for regional disaster mitigation planning.

Key Word: ERA5; Imputation; Statistical Downscaling; Random Forest; XGBoost

Abstrak

Perubahan iklim meningkatkan risiko bencana hidrometeorologi di Maluku Utara, sementara kekosongan data curah hujan BMKG dan bias spasial ERA5 menghambat analisis risiko presisi. Penelitian ini menerapkan Statistical Downscaling berbasis titik menggunakan *Random Forest* (RF) dan *Extreme Gradient Boosting* (XGBoost), dengan presipitasi ERA5 sebagai prediktor tunggal, untuk mengimputasi kekosongan data curah hujan di empat stasiun. XGBoost menghasilkan MSE dan RMSE lebih rendah di tiga stasiun (Sultan Babullah: MSE 1017,29, RMSE 31,89, MAE 12,31, R^2 0,0351; Gamar Malamo: MSE 513,96, RMSE 22,67, MAE 9,91, R^2 -0,0141; Oesman Sadik: MSE 163,59, RMSE 12,79, MAE 6,27, R^2 0,0369), dengan akurasi berbasis NRMSE 83,89 – 87,56%, meski RF tetap mencatat MAE sedikit lebih rendah. RF terpilih di Emalamo (MSE 163,75, RMSE 12,79, MAE 5,62, R^2 -0,0517, akurasi 85,78%). Model berhasil mengimputasi 298 hari data kosong, meski prediktor univariat underestimate curah hujan di atas 40 mm. Basis data historis kontinu menjadi landasan saintifik bagi mitigasi bencana daerah.

Kata kunci: ERA5; Imputasi; Statistical Downscaling; Random Forest; XGBoost

1. Pendahuluan

Perubahan iklim global saat ini telah menjadi tantangan nyata yang mengubah pola cuaca secara drastis, khususnya di wilayah kepulauan tropis. Di Indonesia, fenomena ini ditandai dengan pergeseran awal musim dan peningkatan intensitas curah hujan ekstrem yang semakin sulit diprediksi. Studi sebelumnya mengonfirmasi bahwa pola curah hujan di berbagai wilayah Indonesia telah mengalami perubahan signifikan, yang berimplikasi langsung pada peningkatan risiko bencana hidrometeorologi seperti banjir, tanah longsor, dan angin puting beliung [1]. Provinsi Maluku Utara berada pada posisi yang sangat rentan terhadap perubahan ini karena karakteristik topografinya yang berupa kepulauan dengan kontur pegunungan, sehingga menyebabkan variasi cuaca lokal yang ekstrem antar pulau. Kejadian bencana banjir bandang yang melanda Kelurahan Rua, Kota Ternate pada tahun 2024 [2] menjadi bukti krusial bahwa wilayah ini membutuhkan informasi iklim yang lebih presisi untuk mitigasi bencana.

Upaya pemetaan risiko iklim di Maluku Utara sering terhambat oleh masalah klasik, yaitu keterbatasan dan kekosongan data historis yang kontinu. Data observasi lapangan dari stasiun Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) merupakan acuan kebenaran (*ground truth*) yang paling akurat, namun seringkali mengalami kekosongan data akibat kendala teknis operasional di daerah. Untuk mengatasi kekosongan tersebut, pemanfaatan data reanalisis ERA5 menjadi solusi alternatif karena mampu menyediakan data rekaman historis sirkulasi atmosfer tanpa jeda sejak tahun 1940 dengan resolusi spasial yang relatif detail (~31 km). Akan tetapi, pengembang ERA5 mencatat bahwa data global tersebut masih mengandung bias spasial dan terkadang menghasilkan nilai yang kurang sesuai jika dihadapkan pada wilayah dengan topografi yang rumit [3]. Hal ini turut dikonfirmasi oleh studi-studi terkini yang menegaskan bahwa data ERA5 seringkali gagal menangkap variasi iklim lokal secara presisi pada wilayah heterogen [4], [5]. Penerapan metode koreksi bias dalam *statistical downscaling* ini terbukti mampu meningkatkan akurasi data curah hujan di wilayah dengan karakteristik geografis pegunungan yang rumit [6]. Kondisi inilah yang menjadi objek permasalahan utama dalam penelitian ini: bagaimana merekonstruksi data curah hujan stasiun yang hilang dengan tetap mengoreksi bias spasial data *grid* ERA5 pada wilayah kepulauan bertopografi ekstrem seperti Maluku Utara.

Terkait imputasi curah hujan secara spesifik, Sukhna *et al.* [7] dan Cahyaning *et al.* [8] telah meneliti pemulihan data iklim yang hilang menggunakan konsep *Support Vector Regression* (SVR) dan berhasil merekonstruksi nilai observasi pada variabel temperatur maupun curah hujan. Miftahurrohman *et al.* [9] memperkuat temuan tersebut dengan membuktikan bahwa metode *ensemble learning* (*Random Forest*) memberikan hasil koreksi bias yang lebih baik dibandingkan metode *non-ensemble*. Pada ranah komparasi algoritma, Mursianto *et al.* [10] serta Sibuea & Maulana [11] mengeksplorasi penggunaan *ensemble tree* (XGBoost dan RF) yang terbukti andal menangani data cuaca berderau tinggi. Meskipun demikian, terdapat *gap* yang masih tersisa: penelitian imputasi terdahulu didominasi oleh algoritma SVR dan umumnya diterapkan pada wilayah daratan berskala luas. Belum ada penelitian yang memfokuskan pengujian keandalan metode *ensemble learning* untuk merekonstruksi deret waktu iklim di ekosistem kepulauan bertopografi ekstrem yang sangat rentan terhadap efek penghalusan spasial (*spatial smoothing*).

Untuk menyelesaikan *gap* tersebut, penelitian ini mengusulkan konsep solusi berupa penerapan *point-based Statistical Downscaling* menggunakan algoritma XGBoost dan RF guna mengimputasi kekosongan data curah hujan di empat stasiun BMKG Maluku Utara. Solusi ini dipandang efektif karena pendekatan statistik konvensional terbukti kurang mumpuni menangkap pola hujan tropis yang *non-linear*, sedangkan *Machine Learning* menunjukkan akurasi lebih tinggi dalam memprediksi hujan [12], termasuk saat diintegrasikan ke model sirkulasi iklim untuk memproses persamaan fisika kompleks secara kompetitif [13]. Penelitian global juga membuktikan bahwa *Statistical Downscaling* berbasis ML efektif mereduksi bias ERA5 dan merekonstruksi presipitasi beresolusi tinggi pada wilayah bertopografi ekstrem [14], [15], [16] sementara algoritma *ensemble tree* secara teoretis unggul menangani data cuaca berderau tinggi dan berdistribusi tidak normal [17], [18], [19], [20]. Integrasi teknik *downscaling* presipitasi dengan algoritma *ensemble* ini juga telah divalidasi mampu mereduksi bias data satelit dan meningkatkan akurasi prediksi secara signifikan [21]. Secara spesifik, penelitian ini bertujuan mengevaluasi dan membandingkan kinerja RF dan XGBoost dalam melakukan *point-based downscaling* data ERA5, serta mengaplikasikan model dengan galat terendah pada setiap stasiun untuk mengimputasi data curah hujan yang hilang di empat stasiun UPT BMKG Maluku Utara. *State of the art* (kebaruan) dari riset ini terletak pada adaptasi fungsi transfer *ensemble tree* yang didesain menggunakan presipitasi ERA5 sebagai prediktor tunggal (univariat) untuk memodelkan dan mengimputasi data iklim historis khusus di kepulauan bertopografi ekstrem. Hasil luaran dari model komputasi terpilih diharapkan mampu menghasilkan basis data iklim yang kontinu dan minim bias sebagai landasan saintifik bagi pemerintah daerah dalam merencanakan kebijakan tata ruang dan mitigasi bencana.

2. Metodologi

Penelitian ini menggunakan pendekatan komputasi terapan yang difokuskan pada pengujian kinerja metode *Statistical Downscaling* berbasis titik (*point-based*). Tahapan penelitian dirancang untuk mengatasi kesenjangan resolusi spasial antara model iklim global dan stasiun pengamatan permukaan, yang terbagi ke dalam empat fase utama: pengumpulan bahan dan parameter data, prapemrosesan dan ekstraksi fitur, pembangunan model *Machine Learning*, serta validasi kinerja dan imputasi data.

2.1 Bahan dan Parameter Data

Eksperimen pemodelan ini menggunakan parameter data dalam deret waktu harian yang ditarik secara spesifik pada periode dua tahun kalender (1 Mei 2024 hingga 30 April 2026). Variabel prediktor (masukan) diekstrak dari luaran reanalisis sirkulasi atmosfer global ERA5 (parameter *total precipitation*). Resolusi spasial ERA5 yang digunakan adalah $0.25^\circ \times 0.25^\circ$ (sekitar 31 km x 31 km). Variabel target atau *predictand* (keluaran) merupakan data observasi curah hujan permukaan yang bersumber dari empat stasiun UPT BMKG di Provinsi Maluku Utara. Dalam proses penambangan dan komputasi datanya, variabel *total precipitation* (presipitasi) dari grid ERA5 disetarakan derajatnya secara langsung dengan curah hujan harian dari instrumen BMKG. Rasionalisasi penyetaraan ini didasarkan pada karakteristik klimatologi wilayah Indonesia, secara khusus Kepulauan Maluku Utara, yang berada tepat di garis khatulistiwa beriklim tropis. Di area geografis tersebut, fenomena presipitasi yang jatuh ke permukaan laut maupun darat seutuhnya terjadi dalam wujud curah hujan cair, tanpa adanya fase presipitasi kristal padat seperti salju. Untuk mendukung implementasi algoritma berbasis lokasi (*point-based*), detail koordinat geografis dari setiap stasiun BMKG disajikan pada Tabel 1.

Tabel 1. Daftar Koordinat Stasiun UPT BMKG Maluku Utara

Nama Stasiun	Latitude	Longitude
Sultan Babullah	0.83330	127.38000
Gamar Malamo	1.84092	127.79050
Oesman Sadik	-0.63784	127.50000
Emalamo	-2.05000	126.00000

2.2 Prosedur Prapemrosesan dan Ekstraksi Fitur

Data mentah yang dikumpulkan diolah melalui fase prapemrosesan untuk menjamin integritas statistik sebelum masuk ke fungsi transfer *Machine Learning*. Langkah awal meliputi pembersihan data observasi BMKG dengan mengidentifikasi kode anomali pencatatan dan merepresentasikan kekosongan data akibat kendala operasional menjadi format yang kompatibel untuk komputasi. Selanjutnya, implementasi *point-based downscaling* dieksekusi dengan menyelaraskan resolusi spasial data grid ERA5 secara eksak ke titik koordinat geografis stasiun BMKG terdekat (*nearest neighbor*). Guna mengakomodasi variabilitas siklus monsun tahunan yang bersifat sirkular, dilakukan rekayasa fitur waktu (*feature engineering*) menggunakan transformasi trigonometri dua dimensi (sinus dan kosinus).

2.3 Pembangunan Algoritma Machine Learning

Data yang telah tersinkronisasi (*overlap*) dibagi secara hierarkis menjadi dua segmen: himpunan data latih (*training data*) sebesar 80% dan himpunan data uji (*testing data*) sebesar 20% tanpa pengacakan urutan (*shuffle*) untuk menjaga validitas informasi temporal. Dua algoritma fungsional, yaitu *Random Forest* (RF) dan *Extreme Gradient Boosting* (XGBoost), dibangun dan dilatih menggunakan porsi data latih guna mengenali pola non-linear curah hujan harian. Pemanfaatan kedua algoritma ini didasarkan pada keandalannya dalam memodelkan variabel iklim dan cuaca yang kompleks [10], [11], [12],[18].

Algoritma Random Forest (RF) bekerja menggunakan prinsip *bagging* (*Bootstrap Aggregating*) dengan membangun B buah pohon keputusan (*Decision Trees*) secara independen dari sampel *bootstrap* data latih. Prediksi akhir curah hujan ($\hat{y}$) diperoleh melalui rata-rata luaran dari seluruh pohon penyusun [10], [18], yang dirumuskan pada Persamaan (1):

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B h_b(\mathbf{x}) \quad (1)$$

Di mana $h_b(\mathbf{x})$ menyatakan fungsi prediksi dari pohon keputusan ke- b berdasarkan vektor fitur prediktor $\mathbf{x}$ (presipitasi ERA5 dan fitur waktu siklikal).

Sementara itu, *Extreme Gradient Boosting* (XGBoost) menerapkan pendekatan *gradient boosting* terintegrasi yang menambahkan pohon keputusan secara sekuensial untuk meminimalkan fungsi kerugian (*loss function*) dari iterasi sebelumnya. Pada iterasi ke- t , fungsi tujuan (*objective function*) $\mathcal{L}^{(t)}$ yang dioptimalkan dirumuskan pada Persamaan (2) [19]:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \quad (2)$$

Dengan term regularisasi $\Omega(f_t)$ untuk mencegah *overfitting* yang didefinisikan sebagai Persamaan (3):

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum w_j^2 \quad (3)$$

Di mana l menyatakan fungsi kerugian galat (*loss function*, misalnya *Mean Squared Error*), y_i dan $\hat{y}_i$ masing-masing adalah observasi dan prediksi curah hujan ke- i , T adalah jumlah daun (*leaves*) pada pohon f_t , w_j adalah bobot dari daun ke- j , serta γ dan λ merupakan hiperparameter regularisasi.

2.4 Validasi Kinerja dan Imputasi Data

Validasi akurasi performa metode komputasi pada data uji diukur secara objektif menggunakan metrik regresi komprehensif [9], [18], [21],[22] yang meliputi *Mean Squared Error* (MSE), *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan Koefisien Determinasi (R^2). Secara sistematis, formula matematis untuk masing-masing metrik evaluasi dirumuskan pada Persamaan (4) hingga (7):

1) *Mean Squared Error* (MSE)

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (4)$$

2) *Root Mean Square Error* (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (5)$$

3) *Mean Absolute Error* (MAE)

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (6)$$

4) *R-squared* (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (7)$$

Di mana y_i menyatakan nilai observasi curah hujan BMKG ke- i , $\hat{y}_i$ menyatakan nilai prediksi model ke- i , $\bar{y}$ adalah rata-rata nilai observasi curah hujan, dan N adalah jumlah sampel data uji. Algoritma yang membuktikan tingkat galat kuadratik terendah dikonfigurasi sebagai model definitif. Model komputasi terbaik ini kemudian diaplikasikan secara dinamis untuk melakukan imputasi, yaitu menyuntikkan nilai prediksi curah hujan ke dalam baris data stasiun BMKG yang mengalami kegagalan pencatatan atau kekosongan data. Hasil akhir dari prosedur ini adalah basis data iklim historis yang kontinu. Sebagai pelengkap interpretasi visual, tingkat akurasi model turut dihitung menggunakan pendekatan *Normalized Root Mean Square Error* (NRMSE) [22], yaitu RMSE yang dinormalisasi terhadap rentang (*range*) nilai observasi aktual pada masing-masing stasiun, kemudian dikonversi ke skala persentase sesuai dengan Persamaan (8) dan (9):

1) *Normalized Root Mean Square Error*

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \quad (8)$$

2) Akurasi (%)

$$\text{Akurasi (\%)} = (1 - NRMSE) \times 100\% \quad (9)$$

Di mana y_{max} dan y_{min} masing-masing menunjukkan nilai observasi curah hujan maksimum dan minimum pada stasiun yang bersangkutan.

3. Hasil dan Pembahasan

Eksperimen pemodelan *Statistical Downscaling* dilakukan pada deret waktu harian periode 1 Mei 2024 hingga 30 April 2026 (730 hari kalender). Hasil pengujian komputasi dibagi ke dalam fase prapemrosesan sampel, evaluasi performa algoritma, serta penyelesaian masalah kekosongan data melalui imputasi model.

3.1 Hasil Pemrosesan dan Rekayasa Fitur

Proses pencocokan (*overlap*) antara observasi BMKG dan data grid ERA5 berhasil mengidentifikasi dan memisahkan data valid untuk pelatihan dari data kosong yang menjadi target imputasi. Stasiun Emalamo tercatat memiliki tingkat kehilangan data tertinggi (15,61%), sedangkan Stasiun Oesman Sadik memiliki ketersediaan paling lengkap (94,25%). Rincian sebaran data kosong pada keempat stasiun pengamatan disajikan pada Tabel 2.

Tabel 2. Distribusi Pembagian Data Valid dan Data Kosong Stasiun

Nama Stasiun	Total Hari	Data Overlap	Data Missing	Persentase Data Hilang (%)
Sultan Babullah	730	661	69	9.45
Gamar Malamo	730	657	73	10
Oesman Sadik	730	688	42	5.75
Emalamo	730	616	114	15.61

Tabel di atas memperlihatkan bahwa jumlah data hilang terbesar terdapat di Stasiun Emalamo, yaitu 114 hari atau sekitar 15,61% dari keseluruhan periode. Sebaliknya, Stasiun Oesman Sadik memiliki kelengkapan data yang paling baik dengan hanya 42 hari yang tidak terekam (5,75%). Total keseluruhan data yang perlu diimputasi dari empat stasiun adalah 298 hari.

Untuk mendukung algoritma dalam memetakan pola cuaca di wilayah tropis yang dikontrol kuat oleh siklus monsun, dilakukan rekayasa fitur waktu siklikal. Fitur ini menerapkan transformasi trigonometri dua dimensi menggunakan pasangan fungsi sinus dan kosinus untuk memetakan posisi bulan ke dalam ruang koordinat lingkaran, sehingga bulan Desember dan Januari memiliki jarak geometris yang berdekatan. Rumus yang digunakan adalah:

1) Sin Bulan

$$Bulan_{sin} = \sin\left(\frac{2\pi \cdot m}{12}\right) \quad (1)$$

2) Cos Bulan

$$Bulan_{cos} = \cos\left(\frac{2\pi \cdot m}{12}\right) \quad (2)$$

Di mana m merupakan urutan bulan kalender dan 12 adalah total periode dalam satu siklus tahunan. Meskipun fitur telah direkayasa, uji korelasi Pearson awal antara prediktor ERA5 dan observasi stasiun menunjukkan koefisien yang rendah (berkisar antara 0,1867 hingga 0,3152). Nilai ini merepresentasikan efek penghalusan spasial (*spatial smoothing*), di mana grid makro ERA5 gagal menangkap puncak intensitas hujan konvektif lokal berdurasi singkat secara presisi.

3.2 Pembangunan Model dan Penyetelan Hiperparameter

Dataset yang tervalidasi dibagi dengan rasio 80:20 untuk data latih dan data uji tanpa pengacakan urutan (*shuffle=False*) guna mencegah kebocoran informasi temporal. Guna mengoptimalkan kinerja model, penyetelan hiperparameter (*hyperparameter tuning*) dijalankan menggunakan metode *RandomizedSearchCV* yang dikonfigurasi untuk meminimalkan nilai RMSE. Konfigurasi optimal yang dihasilkan disajikan pada Tabel 3.

Tabel 3. Konfigurasi Hiperparameter Optimal Model *Machine Learning*

Nama Stasiun	Model/Algoritma	Hyperparameter Tuning	Hyperparameter Terpilih	RMSE
Sultan Babullah	Random Forest	<i>max_depth</i> : [3,20] <i>min_samples_leaf</i> : [1,10] <i>min_samples_split</i> : [2,20]	<i>max_depth</i> : 4, <i>min_samples_leaf</i> : 8, <i>min_samples_split</i> : 13	21,06
	XGBoost	<i>max_depth</i> : [3,20] <i>learning_rate</i> : [0.01, 0.20] <i>subsample</i> : [0.60, 0.40] <i>colsample_bytree</i> : [0.60, 0.40]	<i>max_depth</i> : 9, <i>learning_rate</i> : 0,01929, <i>subsample</i> : 0,6682, <i>colsample_bytree</i> : 0,83696	16,71
Gamar Malamo	Random Forest	<i>max_depth</i> : [3,20] <i>min_samples_leaf</i> : [1,10] <i>min_samples_split</i> : [2,20]	<i>max_depth</i> : 4, <i>min_samples_leaf</i> : 8, <i>min_samples_split</i> : 13	13,65
	XGBoost	<i>max_depth</i> : [3,20] <i>learning_rate</i> : [0.01, 0.20] <i>subsample</i> : [0.60, 0.40] <i>colsample_bytree</i> : [0.60, 0.40]	<i>max_depth</i> : 9, <i>learning_rate</i> : 0,01929, <i>subsample</i> : 0,6682, <i>colsample_bytree</i> : 0,83696	9,95
Oesman Sadik	Random Forest	<i>max_depth</i> : [3,20] <i>min_samples_leaf</i> : [1,10] <i>min_samples_split</i> : [2,20]	<i>max_depth</i> : 4, <i>min_samples_leaf</i> : 8, <i>min_samples_split</i> : 13	16,37
	XGBoost	<i>max_depth</i> : [3,20] <i>learning_rate</i> : [0.01, 0.20] <i>subsample</i> : [0.60, 0.40] <i>colsample_bytree</i> : [0.60, 0.40]	<i>max_depth</i> : 9, <i>learning_rate</i> : 0,01929, <i>subsample</i> : 0,6682, <i>colsample_bytree</i> : 0,83696	12,26
Emalamo	Random Forest	<i>max_depth</i> : [3,20] <i>min_samples_leaf</i> : [1,10]	<i>max_depth</i> : 4, <i>min_samples_leaf</i> : 8, <i>min_samples_split</i> : 13	12,67

Nama Stasiun	Model/Algoritma	Hyperparameter Tuning	Hyperparameter Terpilih	RMSE
		<i>min_samples_split</i> : [2,20]		
	XGBoost	<i>max_depth</i> : [3,20] <i>learning_rate</i> : [0.01, 0.20] <i>subsample</i> : [0.60, 0.40] <i>colsample_bytree</i> : [0.60, 0.40]	<i>max_depth</i> : 7, <i>learning_rate</i> : 0,02271, <i>subsample</i> : 0,8918, <i>colsample_bytree</i> : 0,73235	9,54

Penentuan hiperparameter optimal pada algoritma *Random Forest* (RF) dan *Extreme Gradient Boosting* (XGBoost) dilakukan secara eksklusif pada himpunan data latih (*training data*) menggunakan metode *RandomizedSearchCV*. Metode ini bekerja dengan mengevaluasi berbagai kombinasi parameter secara iteratif untuk menemukan konfigurasi yang menghasilkan nilai galat minimum selama proses pelatihan berlangsung. Tingkat galat pada tahap pelatihan tersebut diukur menggunakan metrik *Root Mean Square Error* (RMSE), yang nilai minimumnya untuk setiap stasiun telah dirangkum pada Tabel 3.

Berdasarkan hasil optimasi tersebut, ketiga stasiun pertama (Sultan Babullah, Gamar Malamo, dan Oesman Sadik) menghasilkan konfigurasi hiperparameter *Random Forest* (RF) dan *Extreme Gradient Boosting* (XGBoost) yang identik. Hal ini terjadi karena pada proses pelatihan, karakteristik distribusi data curah hujan di ketiga stasiun tersebut memiliki kedekatan pola, sehingga algoritma konvergen pada ruang parameter yang sama untuk mencapai RMSE terendah. Sebaliknya, proses optimasi pada himpunan data latih di Stasiun Emalamo menghasilkan konfigurasi algoritma XGBoost yang berbeda (khususnya pada *max_depth*: 7 dan *subsample*: 0,8918). Perbedaan konfigurasi ini menunjukkan bahwa variasi curah hujan historis di Stasiun Emalamo memiliki karakteristik sebaran yang spesifik, sehingga model memerlukan pembatasan kedalaman pohon dan rasio sampel yang berbeda agar dapat menginterpretasi data latih secara optimal tanpa mengalami *overfitting*.

3.3 Evaluasi Kinerja Model Komputasi

Tahap evaluasi kinerja komputasi dilakukan untuk mengukur keandalan model yang telah dilatih dengan mengujinya pada himpunan data uji (*testing data*). Pengujian ini membandingkan hasil prediksi luaran dari algoritma *Random Forest* (RF) dan *Extreme Gradient Boosting* (XGBoost) terhadap nilai observasi aktual curah hujan dari BMKG. Ringkasan komparasi performa dari kedua algoritma pada masing-masing stasiun pengamatan disajikan secara komprehensif pada Tabel 4.

Tabel 4. Evaluasi Kinerja Model Komputasi

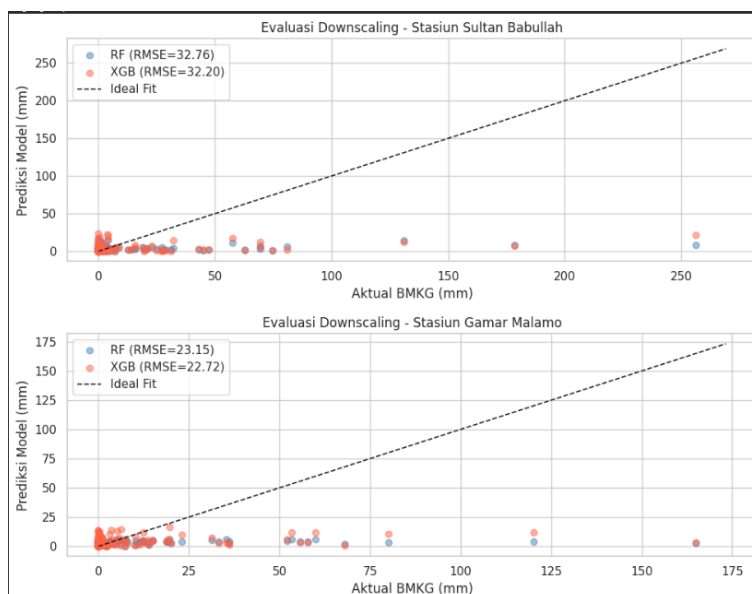
Stasiun	Algoritma	MSE	RMSE	MAE	R2	Akurasi NRMSE (%)	Status
Sultan Babullah	<i>Random Forest</i>	1073,19	32,76	12,17	-0,018	87,22	
	XGBoost	1017,29	31,89	12,31	0,0351	87,56	Model Terpilih
Gamar Malamo	<i>Random Forest</i>	535,99	23,15	9,47	-0,0576	85,97	
	XGBoost	513,96	22,67	9,91	-0,0141	86,26	Model Terpilih

Stasiun	Algoritma	MSE	RMSE	MAE	R ²	Akurasi NRMSE (%)	Status
Oesman Sadik	<i>Random Forest</i>	168,11	12,96	6,18	0,0103	83,67	Model Terpilih
	XGBoost	163,59	12,79	6,27	0,0369	83,89	
Emalamo	<i>Random Forest</i>	163,75	12,79	5,62	-0,0517	85,78	Model Terpilih
	XGBoost	170,54	13,05	6,19	-0,0953	85,49	

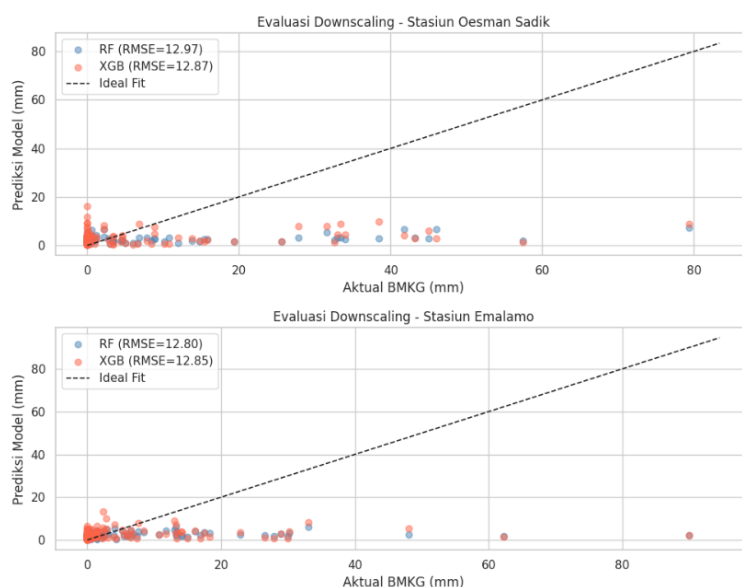
Kinerja komputasi dari kedua algoritma dievaluasi menggunakan empat metrik utama, yaitu *Mean Squared Error* (MSE), *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan Koefisien Determinasi (R^2). Berdasarkan metrik galat kuadrat, algoritma XGBoost secara konsisten menghasilkan nilai MSE dan RMSE yang lebih rendah di tiga stasiun pengamatan (Sultan Babullah, Gamar Malamo, dan Oesman Sadik), menunjukkan bahwa mekanisme iteratif *gradient boosting* pada XGBoost mampu beradaptasi lebih baik dalam meminimalkan galat kuadrat antara nilai prediksi dan observasi aktual curah hujan di ketiga lokasi tersebut. Meskipun demikian, pada metrik galat absolut (MAE), algoritma *Random Forest* (RF) justru secara konsisten mencatatkan nilai yang sedikit lebih rendah dibandingkan XGBoost pada ketiga stasiun yang sama (Sultan Babullah: RF 12,17 vs XGBoost 12,31; Gamar Malamo: RF 9,47 vs XGBoost 9,91; Oesman Sadik: RF 6,18 vs XGBoost 6,27). Pola ini mengindikasikan bahwa meskipun XGBoost lebih unggul dalam menekan galat kuadrat yang sensitif terhadap nilai ekstrem, RF tetap memberikan estimasi titik yang secara rata-rata sedikit lebih dekat dengan observasi aktual pada skala harian.

Sebaliknya, pada Stasiun Emalamo, algoritma Random Forest (RF) menghasilkan nilai galat yang lebih rendah di semua parameter pengukuran (MSE: 163,75; RMSE: 12,79; MAE: 5,62). Pendekatan *bagging* paralel pada algoritma RF menunjukkan stabilitas yang lebih baik dalam memodelkan sebaran data curah hujan dengan karakteristik spesifik di stasiun tersebut dibandingkan XGBoost.

Sementara itu, pengukuran pada Koefisien Determinasi (R^2) menunjukkan hasil yang bervariasi. Model terpilih di Stasiun Gamar Malamo (XGBoost, $R^2=-0,0141$) dan Stasiun Emalamo (RF, $R^2=-0,0517$) menghasilkan nilai R^2 negatif, sedangkan model terpilih di Stasiun Sultan Babullah (XGBoost, $R^2=0,0351$) dan Stasiun Oesman Sadik (XGBoost, $R^2=0,0369$) justru menghasilkan nilai R^2 positif meskipun sangat kecil dan mendekati nol. Selain metrik regresi konvensional, Tabel 4 turut menyajikan tingkat akurasi berbasis *Normalized RMSE* (NRMSE) untuk mempermudah interpretasi kinerja model secara persentase. Hasil pengukuran menunjukkan tingkat akurasi yang relatif tinggi dan konsisten di seluruh stasiun, berkisar antara 83,67% (RF, Oesman Sadik) hingga 87,56% (XGBoost, Sultan Babullah). Pola peringkat akurasi ini sejalan dengan hasil pada metrik MSE dan RMSE, di mana model dengan galat kuadrat terendah pada suatu stasiun juga secara konsisten menghasilkan persentase akurasi tertinggi di stasiun yang sama. Nilai akurasi inilah yang menjadi acuan visual pada diagram pencar (*scatter plot*) di Gambar 1 dan Gambar 2, yang menampilkan persentase akurasi RF dan XGBoost secara berdampingan untuk masing-masing stasiun pengamatan.



Gambar 1. Scatter Plot Evaluasi *Downscaling* Stasiun Sultan Babullah dan Gamar Malamo



Gambar 2. Scatter Plot Evaluasi *Downscaling* Stasiun Oesman Sadik dan Emalamo

Visualisasi pada Gambar 1 dan Gambar 2 mengonfirmasi penyebab anomali nilai Koefisien Determinasi (R^2) yang negatif. Pada diagram pencar tersebut, terlihat bahwa algoritma RF maupun XGBoost mampu memetakan prediksi hari kering (0 mm) dan curah hujan berintensitas ringan dengan sebaran titik yang mendekati garis ideal (*ideal fit*), dibuktikan dengan nilai galat absolut (MAE) yang relatif rendah secara keseluruhan (berkisar antara 5,62 hingga 12,31 mm). Namun, ketika observasi aktual menunjukkan curah hujan di atas 40 mm, sebaran titik prediksi dari kedua model cenderung mendatar secara horizontal. Kondisi *underestimation* pada rentang presipitasi tinggi inilah yang memengaruhi performa evaluasi metrik. Fenomena ini merupakan konsekuensi empiris dari penggunaan prediktor tunggal (univariat) berbasis data makro ERA5, di mana variabel tersebut memiliki keterbatasan dalam merepresentasikan presipitasi mesoskala secara presisi, sehingga model kehilangan fitur krusial untuk memetakan puncak intensitas hujan lokal.

3.4 Hasil Rekonstruksi dan Imputasi Data

Implementasi tahap akhir difokuskan pada penyelesaian masalah utama penelitian, yaitu merekonstruksi data iklim yang terputus. Model terbaik dari hasil pengujian (XGBoost untuk Stasiun Sultan Babullah, Gamar Malamo, dan Oesman Sadik, serta RF untuk Stasiun Emalamo) diaplikasikan untuk melakukan fungsi imputasi. Tabel 5 menyajikan cuplikan empiris representasi proses penyuntikan nilai prediksi komputasi ke dalam pangkalan data observasi BMKG.

Tabel 5. Cuplikan Historis Rekonstruksi (*Before-After*) Imputasi Data

Tanggal	Nama Stasiun	Aktual BMKG (Sebelum)	Luaran Model (Sesudah)
2024-05-13	Sultan Babullah	NaN (Kosong)	2.91 mm
2024-05-08	Gamar Malamo	NaN (Kosong)	1.54 mm
2024-05-19	Oesman Sadik	NaN (Kosong)	1.15 mm
2024-05-02	Emalamo	NaN (Kosong)	3.43 mm

Proses inferensi komputasi ini berhasil merekonstruksi 298 baris hari pengamatan yang kosong secara penuh. Luaran ini memastikan terbentuknya basis data iklim historis yang kontinu untuk Provinsi Maluku Utara, yang luaran datanya dapat diakses secara terbuka melalui tautan: "https://bit.ly/data_bmkg_imputasi"

3.5 Pembahasan

Penelitian ini telah berhasil menjawab tujuan awal yang ditetapkan, yakni menguji kinerja *Statistical Downscaling* berbasis *Machine Learning* dan mengimputasi kekosongan data curah hujan historis di Provinsi Maluku Utara. Hasil komputasi menunjukkan bahwa algoritma *Extreme Gradient Boosting* (XGBoost) dominan dengan galat terendah di tiga stasiun, *sementara Random Forest* (RF) lebih stabil di satu stasiun. Keberhasilan model dalam mengimputasi 298 hari data kosong membuktikan bahwa algoritma *ensemble tree* dapat diandalkan sebagai instrumen rekonstruksi data iklim yang terputus, sehingga mengatasi kendala teknis operasional di stasiun observasi permukaan. Capaian ini secara langsung menjawab permasalahan yang diangkat pada bagian Pendahuluan, yaitu minimnya kontinuitas data curah hujan BMKG akibat kendala teknis di lapangan, sehingga basis data hasil imputasi ini dapat segera dimanfaatkan sebagai masukan analisis risiko bencana hidrometeorologi di Provinsi Maluku Utara tanpa harus menunggu perbaikan infrastruktur pemantauan secara fisik.

Jika ditinjau dari landasan konsep komputasi ilmiah, algoritma XGBoost mengungguli RF pada metrik galat kuadratik di mayoritas stasiun karena mekanisme *gradient boosting* bekerja secara sekuensial dalam meminimalkan *residual error* pada setiap iterasi pembentukan pohon keputusan. Pendekatan ini secara teoritis lebih sensitif dalam menangkap pola *non-linear* dari fluktuasi cuaca ekstrem dibandingkan dengan mekanisme *bagging* paralel pada RF [23]. Namun, keunggulan ini tidak bersifat mutlak di seluruh metrik. Seperti yang telah dijabarkan pada sub-bab 3.3, RF justru secara konsisten mencatatkan MAE yang sedikit lebih rendah dibandingkan XGBoost pada ketiga stasiun yang sama. Pola ini dapat dijelaskan secara teoretis bahwa mekanisme *bagging* paralel pada RF menghasilkan estimasi rata-rata (*averaging*) dari banyak pohon independen yang cenderung lebih *robust* terhadap galat titik-demi-titik, sehingga menghasilkan deviasi absolut rata-rata yang lebih rendah. Sebaliknya, XGBoost yang dioptimalkan untuk meminimalkan galat kuadratik secara sekuensial cenderung lebih sensitif terhadap nilai ekstrem (*outlier*), sehingga unggul pada MSE dan RMSE namun tidak selalu unggul pada MAE. Dengan demikian, pemilihan XGBoost sebagai model terpilih pada penelitian ini didasarkan pada prioritas meminimalkan galat kuadratik (RMSE) sebagai kriteria utama optimasi hiperparameter, bukan karena keunggulannya bersifat mutlak pada seluruh metrik evaluasi. Karakteristik topografi Provinsi Maluku Utara yang heterogen menghasilkan derau (*noise*) spasial yang tinggi pada data ERA5, dan XGBoost terbukti lebih adaptif dalam memproses kompleksitas fungsi transfer tersebut. Keunggulan ini sejalan dengan konsep dasar algoritma yang diperkenalkan oleh Chen dan Guestrin [19], di mana mekanisme regularisasi (*regularization*) bawaan pada fungsi objektif XGBoost membantu menekan risiko *overfitting* sekaligus mempertahankan sensitivitas terhadap pola *non-linear*. Sebaliknya, performa *Random Forest* (RF) yang relatif lebih unggul di Stasiun Emalamo diduga terkait dengan konfigurasi kedalaman pohon (*max_depth*) yang dangkal dan homogen antar pohon penyusun, sehingga pendekatan *bagging* paralel pada RF justru lebih stabil ketika distribusi curah hujan di stasiun tersebut

menunjukkan variabilitas historis yang lebih ekstrem dan rentan terhadap *overfitting* apabila dimodelkan secara sekuensial.

Temuan dalam penelitian ini memperkuat hasil studi sebelumnya yang menyatakan bahwa XGBoost memiliki kinerja evaluasi yang lebih baik dalam memprediksi data cuaca dibandingkan algoritma konvensional lainnya [24]. Meskipun demikian, penelitian ini juga menyajikan temuan baru bahwa pada wilayah kepulauan dengan efek penghalusan spasial (*spatial smoothing*) yang ekstrem, penggunaan algoritma *ensemble* terkuat sekalipun akan menghadapi batas jenuh (*underestimation*) jika hanya bergantung pada presipitasi ERA5 sebagai prediktor univariat. Hal ini membedakan penelitian ini dari riset terdahulu yang umumnya berfokus pada wilayah daratan luas dengan kondisi iklim yang lebih homogen [25]. Perbedaan ini juga tampak jika dibandingkan dengan riset imputasi curah hujan berbasis *Support Vector Regression* oleh Sukhna et al. [7] dan Cahyaning et al. [8]. Kedua penelitian tersebut berhasil merekonstruksi data iklim yang hilang di wilayah dataran, namun tidak menguji ketahanan model terhadap efek topografi kepulauan yang ekstrem seperti pada penelitian ini. Sejalan dengan Sangaji dan Sutabri [12] yang juga membandingkan XGBoost dan RF pada konteks curah hujan di Indonesia, penelitian ini memperkuat konsistensi keunggulan XGBoost lintas wilayah geografis, sekaligus menegaskan bahwa faktor topografi lokal turut menentukan algoritma mana yang paling sesuai diterapkan pada suatu stasiun pengamatan.

Secara praktis maupun teoretis, penelitian ini memberikan kontribusi yang signifikan bagi disiplin Sistem Informasi dan Meteorologi Komputasi. Kajian ini menawarkan kerangka kerja alternatif berbasis pembelajaran mesin untuk melakukan imputasi pada deret waktu meteorologi, yang dapat dijadikan rujukan (*baseline*) bagi peneliti lain dalam menangani masalah kekosongan data historis di wilayah dengan topografi serupa. Luaran data kontinu yang dihasilkan juga menjadi aset berharga bagi pemangku kepentingan dalam analisis risiko kebencanaan hidrometeorologi. Secara lebih konkret, basis data hasil imputasi yang telah dipublikasikan secara terbuka dapat langsung dimanfaatkan oleh Pemerintah Daerah Provinsi Maluku Utara maupun BMKG setempat sebagai masukan kuantitatif dalam penyusunan peta rawan bencana dan penentuan prioritas wilayah mitigasi, tanpa perlu menunggu kelengkapan data observasi lapangan yang seringkali memerlukan waktu perbaikan infrastruktur bertahun-tahun.

Meski demikian, penelitian ini memiliki keterbatasan utama pada absennya prediktor mesoskala. Pemodelan univariat gagal menangkap puncak curah hujan di atas 40 mm akibat ketidakmampuan data makro dalam merepresentasikan awan konvektif lokal secara presisi. Oleh karena itu, terdapat peluang besar bagi penelitian selanjutnya untuk mengoptimalkan kelemahan tersebut. Rekomendasi pengembangan mencakup peralihan ke pemodelan multivariat yang mengintegrasikan parameter seperti suhu puncak awan, kelembapan udara, dan kecepatan angin [26]. Selain itu, integrasi model hibrida *deep learning* sekuensial seperti *Long Short-Term Memory* (LSTM) dapat dieksplorasi untuk merekam dependensi pola monsun historis secara lebih akurat. Selain keterbatasan prediktor, ukuran periode data historis yang digunakan (dua tahun kalender) relatif singkat untuk menangkap variabilitas antar-tahunan seperti fenomena *El Nino-Southern Oscillation* (ENSO) dan *Indian Ocean Dipole* (IOD) yang diketahui turut memengaruhi anomali curah hujan di wilayah timur Indonesia [26]. Oleh sebab itu, replikasi metode pada rentang data yang lebih panjang serta validasi silang (*cross-validation*) antar musim akan memperkuat generalisasi model sebelum diadopsi secara operasional oleh instansi terkait.

4. Simpulan

Penelitian ini menyimpulkan bahwa metode *Statistical Downscaling* berbasis *Machine Learning* efektif untuk mengatasi kesenjangan resolusi dan kekosongan data curah hujan historis di wilayah kepulauan Provinsi Maluku Utara. Berdasarkan hasil evaluasi kinerja komputasi, algoritma *Extreme Gradient Boosting* (XGBoost) menunjukkan performa terbaik pada metrik galat kuadratik (MSE dan RMSE) di tiga stasiun pengamatan (Sultan Babullah, Gamar Malamo, dan Oesman Sadik), sementara algoritma *Random Forest* (RF) tampil lebih stabil di Stasiun Emalamo serta secara konsisten mencatatkan MAE yang sedikit lebih rendah di ketiga stasiun lainnya, dengan tingkat akurasi keseluruhan berbasis NRMSE berkisar antara 83,67%–87,56% di seluruh stasiun. Proses inferensi model secara keseluruhan telah berhasil mencapai tujuan utamanya dengan merekonstruksi dan mengimputasi 298 baris hari data yang terputus, sehingga menghasilkan basis data iklim yang kontinu. Basis data hasil imputasi ini memberikan kontribusi ganda, yaitu sebagai kerangka kerja rujukan (*baseline*) bagi penelitian imputasi data iklim di wilayah bertopografi serupa, sekaligus sebagai basis kuantitatif bagi Pemerintah Daerah Provinsi

Maluku Utara dan BMKG setempat dalam perencanaan mitigasi bencana hidrometeorologi. Meskipun pemanfaatan presipitasi ERA5 sebagai prediktor univariat memiliki batas jenuh (*underestimation*) pada rentang hujan ekstrem, luaran data dari pemodelan ini tetap relevan sebagai rujukan data historis. Sebagai tindak lanjut, penelitian ke depan direkomendasikan untuk beralih menggunakan pendekatan pemodelan multivariat dan mengeksplorasi arsitektur *deep learning* untuk meningkatkan akurasi pengenalan pola cuaca lokal.

Daftar Referensi

- [1] S. Suhadi, F. Mabruroh, A. Wiyanto, and I. Ikra, "Analisis Fenomena Perubahan Iklim Terhadap Curah Hujan Ekstrem," *Opt. J. Pendidik. Fis.*, vol. 7, no. 1, pp. 94–100, 2023, doi: 10.37478/optika.v7i1.2738.
- [2] Mongabay, "Tragedi Banjir Bandang Ternate: Puluhan Korban Jiwa, Pencarian Masih Berlangsung," Mongabay Indonesia. [Online]. Available: <https://mongabay.co.id/2024/08/26/tragedi-banjir-bandang-ternate-puluhan-korban-jiwa-pencarian-masih-berlangsung/>
- [3] H. Hersbach *et al.*, "The ERA5 global reanalysis," *Q. J. R. Meteorol. Soc.*, vol. 146, no. 730, pp. 1999–2049, 2020, doi: 10.1002/qj.3803.
- [4] W. Hu, Y. Scholz, M. Yeligeti, L. von Bremen, and Y. Deng, "Downscaling ERA5 wind speed data: a machine learning approach considering topographic influences," *Environ. Res. Lett.*, vol. 18, no. 9, p. 094007, Sep. 2023, doi: 10.1088/1748-9326/aceb0a.
- [5] A. Chakri *et al.*, "Spatial Bias Correction of ERA5_Ag Reanalysis Precipitation Using Machine Learning Models in Semi-Arid Region of Morocco," *Atmosphere (Basel)*, vol. 16, no. 11, pp. 1–35, 2025, doi: 10.3390/atmos16111234.
- [6] R. Rahimi, H. Tavakol-Davani, and M. Nasseri, "An Uncertainty-Based Regional Comparative Analysis on the Performance of Different Bias Correction Methods in Statistical Downscaling of Precipitation," *Water Resour. Manag.*, vol. 35, no. 8, pp. 2503–2518, 2021, doi: 10.1007/s11269-021-02844-0.
- [7] I. K. Sukhna, B. Miftahurrohman, C. Wulandari, and P. Amelia, "Imputasi Data Temperatur Maksimum Menggunakan Metode Support Vector Regression," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 10, no. 2, pp. 171–185, 2025, doi: 10.14421/jiska.2025.10.2.171-185.
- [8] A. Cahyaning, B. Miftahurrohman, G. F. Prassida, and T. Tikno, "Filling the Precipitation Gaps: Accurate Imputation With Support Vector Regression in North Sulawesi," *Var. J. Stat. Its Appl.*, vol. 6, no. 2, pp. 183–194, 2024, doi: 10.30598/variancevol6iss2page183-194.
- [9] B. Miftahurrohman, H. Kuswanto, D. S. Pambudi, F. Fauzi, and F. Atmaja, "Assessment of the Support Vector Regression and Random Forest Algorithms in the Bias Correction Process on Temperatures," *Procedia Comput. Sci.*, vol. 234, pp. 637–644, 2024, doi: 10.1016/j.procs.2024.03.049.
- [10] G. A. Mursianto, I. M. Fali, M. Irfan, T. Sakinah, and D. S. Prasvita, "Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan," *J. Senamika*, vol. 2, no. 2, pp. 41–50, 2021.
- [11] R. U. Sibuea and H. Maulana, "Comparison of Random Forest and XGBOOST Methods on Weather in North Sumatera," *Tsabit J. Comput.*, vol. 2, no. 1, pp. 25–34, 2025, doi: 10.56211/tsabit44.
- [12] D. Sangaji and T. Sutabri, "Analisis XGBoost dan Random Forest untuk Prediksi Curah Hujan dalam Mendukung Mitigasi Karhutla," *J. Pustaka AI (Pusat Akses Kaji. Teknol. Artif. Intell.)*, vol. 5, no. 1, pp. 13–18, 2025, doi: 10.55382/jurnalpustakaai.v5i1.905.
- [13] D. Kochkov *et al.*, "Neural general circulation models for weather and climate," *Nature*, vol. 632, no. 8027, pp. 1060–1066, 2024, doi: 10.1038/s41586-024-07744-y.
- [14] H. Sun *et al.*, "Corrected ERA5 Precipitation by Machine Learning Significantly Improved Flow Simulations for the Third Pole Basins," *J. Hydrometeorol.*, vol. 23, no. 10, pp. 1663–1679, 2022, doi: 10.1175/JHM-D-22-0015.1.
- [15] J. Baño-Medina, R. Manzanar, and J. M. Gutierrez, "Configuration and intercomparison of deep learning neural models for statistical downscaling," *Geosci. Model Dev.*, vol. 13, no. 4, pp. 2109–2124, 2020, doi: 10.5194/gmd-13-2109-2020.
- [16] B. Miftahurrohman, I. A. Cholilie, S. U. Wijaya, F. Atmaja, T. Bariyah, and C. Wulandari, "Future Growing Seasons: Bias Correction with SVR and QDM for Indonesian Temperature Projection under RCP 2.6 and RCP 8.5," vol. 2, no. 2, pp. 306–312, 2024, doi:

- <https://doi.org/10.59796/jcst.V15N2.2025.100> Future.
- [17] M. J. Rahaman, "A Comprehensive Review to Understand the Definitions, Advantages, Disadvantages and Applications of Machine Learning Algorithms," *Int. J. Comput. Appl.*, vol. 186, no. 31, pp. 43–47, 2024, doi: 10.5120/ijca2024923868.
 - [18] A. Syahreza, N. K. Ningrum, and M. A. Syahrazy, "Perbandingan Kinerja Model Prediksi Cuaca: Random Forest, Support Vector Regression, dan XGBoost," *Edumatic J. Pendidik. Inform.*, vol. 8, no. 2, pp. 526–534, 2024, doi: 10.29408/edumatic.v8i2.27640.
 - [19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Aug, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
 - [20] U. N. Christanto, B. Miftahurrohman, T. Bariyah, H. Kuswanto, and N. Faria, "Cluster-Based Machine Learning Approaches for Predicting Daily Maximum Temperatures in Indonesia Under Climate Change," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 11, no. 1, pp. 236–249, 2025, doi: 10.33480/jitk.v11i1.6749.
 - [21] F. Atmaja, B. Miftahurrohman, I. A. Cholilie, S. U. Wijaya, N. Haliza, and Y. A. Utami, "Improving Agricultural Productivity in Indonesia Using Xgboost Precipitation Prediction with Downscaling and Bias Correction," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, Jul. 2024, pp. 485–490. doi: 10.1109/ICICoS62600.2024.10636822.
 - [22] J. Vega-Durán, B. Escalante-Castro, F. A. Canales, G. J. Acuña, and B. Kaźmierczak, "Evaluation of Areal Monthly Average Precipitation Estimates from MERRA2 and ERA5 Reanalysis in a Colombian Caribbean Basin," *Atmosphere (Basel)*, vol. 12, no. 11, p. 1430, Oct. 2021, doi: 10.3390/atmos12111430.
 - [23] Sartika Lina Mulani Sitio, Darmawati, and Yuda Samudra, "COMPARATIVE ANALYSIS OF BAGGING AND BOOSTING MODELS IN ENSEMBLE LEARNING FOR GRADUATION PREDICTION," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 11, no. 3, pp. 979–986, Feb. 2026, doi: 10.33480/jitk.v11i3.7579.
 - [24] R. A. Hasyani, Y. M. Rangkuti, E. Simamora, S. Iskandar, and A. Idrus, "Extreme Gradient Boosting for Daily Rainfall Forecasting in Medan City," *JAIEA (Journal Artif. Intell. Eng. Appl.)*, vol. 5, no. 2, pp. 3155–3163, 2026, doi: doi.org/10.59934/jaiea.v5i2.2136.
 - [25] A. Safril and A. Ulfiana, "Prediksi Curah Hujan Bulanan Menggunakan Metode Statistical Downscaling di Pulau Jawa dengan Pemilihan Prediktor Berdasarkan Transpor Uap Air," *J. Geogr.*, vol. 11, no. 1, pp. 19–33, Feb. 2019, doi: 10.24114/jg.v11i1.11474.
 - [26] B. K. Cahyono *et al.*, "Leveraging machine learning and open accessed remote sensing data for precise rainfall forecasting," *Commun. Sci. Technol.*, vol. 10, no. 1, pp. 135–147, 2025, doi: 10.21924/cst.10.1.2025.1638.