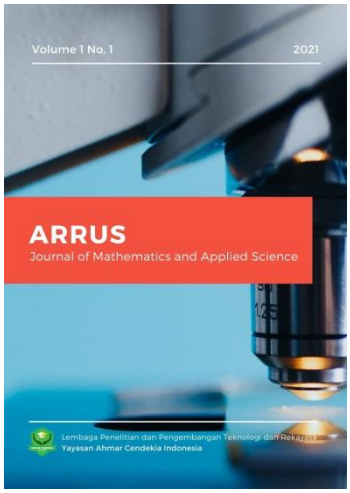
https://doi.org/10.35877/mathscience5607

OPEN ACCESS

ISSN : 2776-7922 (Print) / 2807-3037 (Online)



RESEARCH ARTICLE

Linear Time-Series Models Match Deep Learning in Daily Solar Radiation Forecasting for Tropical Indonesia

Zulkifli Rais[1*] & Muh. Qodri Alfairus[2]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Makassar, South Sulawesi, Indonesia.
[2]Department of Accounting, Faculty of Economics and Business, Universitas Negeri Makassar, Makassar, South Sulawesi, Indonesia.

*Corresponding author: Zulkifli Rais, Department of Statistics, Universitas Negeri Makassar, Makassar, Indonesia

E-mail: zulkifli.rais89@unm.ac.id

Abstract: Comparisons of deep learning forecasts seldom separate the effect of architecture from that of location or test the networks against linear models that use the same inputs. This study compared LSTM, GRU, and BiLSTM networks with two linear time-series models, dynamic harmonic regression with ARMA errors (DHR-ARMA) and an autoregressive model with exogenous inputs (ARX), for one-day-ahead forecasting of daily solar radiation in five Eastern Indonesian cities using NASA POWER data for 2014–2023, with 2023 as the test year. Inference used block-bootstrap confidence intervals, Diebold–Mariano tests with Holm correction, a two-way ANOVA of per-seed RMSE, and Mincer–Zarnowitz and Ljung–Box diagnostics. City explained 99.9% of the variation in RMSE, whereas architecture had no significant effect ($F(2, 60) = 1.50$, $p = 0.231$). None of the 15 comparisons among the networks or the 30 comparisons between networks and linear models remained significant after Holm adjustment; the linear models came within 3.2% of the best network and were best in Kupang. Every model significantly outperformed persistence except in Kupang. The best RMSE ranged from 0.717 to 1.294 kWh/m²/day and followed the day-to-day autocorrelation of radiation. Network forecasts were biased low in four cities, and unbiasedness was rejected in 9 of 15 cases.

Keywords: Diebold–Mariano Test, Forecast Evaluation, Recurrent Neural Network, Solar Radiation, Time Series Analysis

1. Introduction

Indonesia has abundant solar energy resources. A review of Indonesia's renewable energy potential shows that the technical potential of solar photovoltaics (PV) far exceeds that of other renewable sources, and that the official government estimate of 208 GWp is conservative compared with academic and industry estimates in the terawatt range (Langer et al., 2021). Silalahi et al. (2021) estimated that Indonesia's electricity demand of about 9,000 TWh per year in 2050 could be met by PV on roughly 35,000 km² of land, with a mean horizontal irradiation of 4.8 kWh/m²/day.

Eastern Indonesia is a particularly relevant region for PV. Fuel distribution and grid extension are difficult in an archipelago of more than 17,000 islands, and Veldhuis and Reinders (2013) showed that in eastern Indonesia the cost of electricity from grid-connected urban PV was already lower than the generation costs at that time. Cities in the region are also seeking ways

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

to lower their carbon emissions; in Makassar, for example, the contribution of urban green space to CO_2 reduction has been modeled spatially (Ruliana et al., 2025), and rooftop PV offers a complementary route on the supply side. However, PV output depends on solar radiation, which varies from day to day with season, cloud cover, and humidity. As PV penetration increases, accurate radiation forecasts become essential for grid stability and for the scheduling and dispatch of generating units (Antonanzas et al., 2016).

Recurrent deep learning models such as LSTM, GRU, and BiLSTM are increasingly used and compared for solar radiation forecasting, as reviewed in Section 2. From a statistical standpoint, however, such comparisons raise four questions that are seldom addressed together. First, a network should be compared not only with naive benchmarks but also with linear time-series models that use the same information; on a large set of M3 competition series, popular machine learning methods were less accurate than simple statistical methods (Makridakis et al., 2018). Second, differences in RMSE are estimates from a finite test sample and require formal tests of equal predictive accuracy (Diebold & Mariano, 1995; Harvey et al., 1997); when many pairs of models and locations are tested, the family-wise error rate must also be controlled (Holm, 1979). Third, a network trained with different random seeds yields different forecasts, so the variability due to training must be separated from the variability due to the model and the location. Fourth, accuracy measures do not reveal whether forecasts are biased or leave exploitable autocorrelation in their errors, which calls for calibration diagnostics (Mincer & Zarnowitz, 1969; Ljung & Box, 1978).

This study therefore aims (1) to compare LSTM, GRU, and BiLSTM with two linear time-series models, dynamic harmonic regression with ARMA errors (DHR-ARMA) and an autoregressive model with exogenous inputs (ARX), for one-day-ahead forecasting of daily solar radiation in five Eastern Indonesian cities; (2) to decompose the variation in accuracy into city, architecture, and training-seed components with a two-way analysis of variance; (3) to test equal predictive accuracy with the Diebold–Mariano test under Holm correction and to quantify uncertainty with block-bootstrap confidence intervals; (4) to assess forecast calibration; and (5) to translate the forecasts into PV specific energy yield.

2. Literature Review

Research on solar radiation and PV power forecasting has grown rapidly over the past decade (Yang et al., 2018), with machine learning methods widely applied (Voyant et al., 2017). Among deep learning methods, recurrent neural networks (RNNs) are among the most widely used because they capture nonlinear patterns in time series data (Wang et al., 2019). Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) overcome the vanishing-gradient problem of standard RNNs (Bengio et al., 1994) through gating mechanisms; the Gated Recurrent Unit (GRU) (Cho et al., 2014) reduces the LSTM to two gates; and the Bidirectional LSTM (BiLSTM) (Schuster & Paliwal, 1997; Graves & Schmidhuber, 2005) reads the input window in both directions. Qing and Niu (2018) reported that LSTM reduced the RMSE by 18.34% relative to a backpropagation neural network for hourly day-ahead radiation in Cape Verde, Ghimire et al. (2019) and Kumari and Toshniwal (2021) found hybrid CNN-LSTM models superior to single models, Wojtkiewicz et al. (2019) and Brahma and Wadhvani (2020) compared GRU, LSTM, and related variants, and Zameer et al. (2023) found that BiLSTM yielded the lowest error.

Machine learning forecasting is increasingly applied to Indonesian time series, for example, support vector regression (SVR) for the air quality index in Makassar (Rahmat et al., 2023), SVR versus SutteARIMA for coal prices (Ahmar et al., 2025), and a hybrid ARIMA–LSTM for the rupiah–US dollar exchange rate (Rais, Sumarno, et al., 2026). For meteorological variables, gradient boosting and classification-tree models have been used to classify rainfall (Hafid et al., 2025; Rais, Hafid, et al., 2026). For solar radiation, however, studies in Indonesia still mostly apply a single LSTM at a single location, for example, the prediction of global horizontal irradiance to estimate electricity cost savings by Pradipta and Haviana (2025). A comparison of recurrent architectures and linear time-series models across several Eastern

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Indonesian cities with different climates, with multiplicity-adjusted significance tests, has not been reported. Yet this region spans a semi-arid climate (Kupang), a monsoonal rainfall pattern peaking in December–February (Makassar), a local rainfall pattern peaking in mid-year (Ambon), and near-equatorial areas that are relatively wet throughout the year (Manado and Jayapura); the southern monsoonal and the anti-phase Moluccan patterns are two of the three dominant rainfall regions in Indonesia (Aldrian & Susanto, 2003). This climatic contrast makes it possible to ask whether forecast accuracy depends more on the model or on the location.

3. Research Method and Materials

The analysis proceeded in four stages. Daily data for five cities were first described and diagnosed as time series, then modeled with three recurrent networks and four statistical benchmarks, evaluated statistically on the 2023 test year, and finally converted into PV energy yield.

3.1. Study area and data

This study used daily data from 1 January 2014 to 31 December 2023 (3,652 days per city) from NASA POWER, a global meteorological and solar energy data service that is freely accessible through an application programming interface (Sparks, 2018). POWER surface radiation for this period is derived from the CERES SYN1deg satellite product on a 1° × 1° grid, so each city is represented by the grid cell containing its coordinates. Validation of this product against 85 surface stations showed a monthly mean bias in downward shortwave radiation of 3.0 W/m² or 5.7% (Rutan et al., 2015). Data were retrieved for the central coordinates of Kupang (10.17°S, 123.61°E), Makassar (5.15°S, 119.43°E), Ambon (3.70°S, 128.18°E), Manado (1.47°N, 124.84°E), and Jayapura (2.53°S, 140.72°E). The target was the all-sky surface shortwave downward irradiance (ALLSKY_SFC_SW_DWN, kWh/m²/day), and the inputs were this variable together with the clear-sky irradiance (CLRSKY_SFC_SW_DWN, kWh/m²/day), air temperature at 2 m (T2M, °C), relative humidity at 2 m (RH2M, %), wind speed at 2 m (WS2M, m/s), and bias-corrected precipitation (PRECTOTCORR, mm/day). Two calendar features, $\sin(2\pi d/365.25)$ and $\cos(2\pi d/365.25)$, where d is the day of the year, were added to capture the seasonal cycle.

3.2. Data partition and time-series diagnostics

No missing values (code −999) were found in the data for any city. The data were split chronologically into training (2014–2021, n = 2,922 days), validation (2022), and test (2023, n = 365) sets, so that no future information entered model estimation. Each forecast used the 14 days $t-13, \ldots, t$ to predict the radiation on day $t+1$.

Before modeling, the training series of each city was decomposed into a deterministic seasonal component and an anomaly. The seasonal component was a Fourier climatology, Eq. (1), whose number of harmonics K was chosen by the Bayesian information criterion (BIC; Schwarz, 1978):

$$\mu_t = \beta_0 + \sum_{k=1}^{K}\left[\beta_k \cos\left(\frac{2\pi k d_t}{365.25}\right) + \gamma_k \sin\left(\frac{2\pi k d_t}{365.25}\right)\right], \qquad u_t = y_t - \mu_t \tag{1}$$

The anomaly u_t was tested for a unit root with the augmented Dickey–Fuller (ADF) test (Dickey & Fuller, 1979), with a constant and the lag order chosen by the Akaike information criterion (AIC; Akaike, 1974), using the critical values of MacKinnon (1996). Because the ADF null hypothesis is non-stationarity, the KPSS test of level stationarity (Kwiatkowski et al., 1992) was applied as a complement, with a Bartlett kernel and 28 lags. Serial dependence was measured by the autocorrelation function and the Ljung–Box statistic, Eq. (2), where r_k is the lag-k sample autocorrelation (Ljung & Box, 1978):

$$Q(m) = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} \tag{2}$$

3.3. Recurrent neural networks

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

LSTM. The LSTM cell regulates the flow of information through a forget gate (f), an input gate (i), and an output gate (o), as given in Eq. (3):

$$\begin{aligned} f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{3}$$

GRU. The GRU uses an update gate (z) and a reset gate (r) without a separate memory cell. The reset-after variant, the Keras default, was used (Eq. 4):

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ \tilde{h}_t &= \tanh(W_h x_t + r_t \odot (U_h h_{t-1} + b'_h) + b_h) \\ h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \end{aligned} \tag{4}$$

BiLSTM. The BiLSTM runs a forward and a backward LSTM over the same 14-day window and concatenates the forward and backward hidden states; because the backward pass also stays within the historical window, no information after day t is used.

The three networks had equivalent structures: one recurrent layer with 64 units (64 per direction for BiLSTM), a dropout rate of 0.2, and one dense output layer. Inputs were min–max scaled to [0, 1] with training parameters. Training minimized the mean squared error with the Adam optimizer (Kingma & Ba, 2015) at a learning rate of 0.001, a batch size of 32, at most 200 epochs, and early stopping with a patience of 20 epochs on the validation loss. Each network was trained with five random seeds, giving 75 trained networks (5 cities × 3 architectures × 5 seeds). The networks were implemented in Python with NumPy following the Keras/TensorFlow 2 defaults (Glorot-uniform input weights, orthogonal recurrent weights, an LSTM forget-gate bias of 1, and Adam parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-7}$); gradients computed by backpropagation through time agreed with finite-difference gradients to a maximum relative error of 3.0×10^{-6}.

3.4. Statistical benchmarks

Four statistical benchmarks were fitted to the 2014–2021 training data and then used, with parameters fixed, to produce one-step-ahead forecasts for 2023. **Persistence** sets tomorrow's radiation equal to today's. **Monthly climatology** uses the training-period mean of the calendar month.

DHR-ARMA. Following the dynamic harmonic regression approach (Hyndman & Athanasopoulos, 2021), the Fourier climatology of Eq. (1) was combined with an ARMA(p, q) model for the anomaly (Box et al., 2015), Eq. (5), where B is the backshift operator and ε_t is white noise:

$$y_t = \mu_t + u_t, \qquad \left(1 - \sum_{i=1}^{p} \phi_i B^i\right) u_t = \left(1 + \sum_{j=1}^{q} \theta_j B^j\right) \varepsilon_t \tag{5}$$

The orders were selected by AIC over $p \le 3$ and $q \le 2$, and the parameters were estimated by conditional sum of squares. The one-step forecast is the Fourier climatology of day $t + 1$ plus the ARMA forecast of the anomaly made on day t.

ARX. To isolate the contribution of nonlinearity, a linear regression was fitted by ordinary least squares on exactly the information available to the networks, namely 14 lags of the six NASA POWER variables listed in Section 3.1 and the calendar features of day t (Eq. 6), giving 87 parameters:

$$\hat{y}_{t+1} = \alpha + \sum_{j=0}^{13} \mathbf{x}_{t-j}^{\top} \boldsymbol{\beta}_j + \delta_1 \sin\left(\frac{2\pi d_t}{365.25}\right) + \delta_2 \cos\left(\frac{2\pi d_t}{365.25}\right) \tag{6}$$

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

3.5. *Statistical evaluation*

Accuracy and uncertainty. Accuracy on the 365 test days was measured in kWh/m²/day by the measures in Eq. (7), where $e_t = y_t - \hat{y}_t$, and by the skill score relative to persistence, SS = 1 − $\text{RMSE}_{\text{model}}/\text{RMSE}_{\text{persistence}}$ (Marquez & Coimbra, 2013). A positive mean error (ME) indicates underestimation.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}, \quad \text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|e_t|, \quad R^2 = 1 - \frac{\sum_t e_t^2}{\sum_t (y_t - \bar{y})^2}, \quad \text{ME} = \frac{1}{n}\sum_{t=1}^{n} e_t \tag{7}$$

Because daily errors are serially correlated, 95% confidence intervals for RMSE and for differences in RMSE were obtained with the circular block bootstrap (Künsch, 1989; Politis & Romano, 1992), using blocks of 7 days ($\approx n^{1/3}$) and 10,000 replicates, and taking percentile intervals. For the networks, the bootstrap, the tests below, and the calibration diagnostics used the run from the seed with the lowest validation loss, so that model selection did not use the test data.

Sources of variation. The test RMSE of the 75 networks was analyzed with a two-way fixed-effects ANOVA with replication (Montgomery, 2017), Eq. (8), where α_i is the effect of city i, β_j the effect of architecture j, $(\alpha\beta)_{ij}$ their interaction, and the residual ε_{ijk} reflects seed-to-seed training variability (k = 1, …, 5):

$$\text{RMSE}_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{8}$$

Effect sizes are reported as the share of the total sum of squares (η^2) and as partial η^2. Homogeneity of variances was checked with the Brown–Forsythe test (Brown & Forsythe, 1974) and normality of residuals with the Shapiro–Wilk test (Shapiro & Wilk, 1965); as a robustness check, the ANOVA was repeated on log RMSE.

Equal predictive accuracy. Pairs of forecasts were compared with the Diebold–Mariano (DM) test (Diebold & Mariano, 1995) on the squared-error loss differential $d_t = e_{1,t}^2 - e_{2,t}^2$, with the small-sample correction of Harvey et al. (1997), Eq. (9); DM is referred to a t distribution with $n - 1$ degrees of freedom, and a negative value means that the first model is more accurate:

$$\text{DM} = \frac{\bar{d}}{\sqrt{\hat{\gamma}_0/n}}, \qquad \text{DM}^* = \text{DM}\sqrt{\frac{n+1-2h+h(h-1)/n}{n}}, \quad h = 1 \tag{9}$$

The tests were organized into five families of 15 tests each (three comparisons in each of five cities): (A) pairs of networks; (B) networks versus persistence; (C) networks versus ARX; (D) networks versus DHR-ARMA; and (E) the two linear models versus persistence and versus each other. Within each family, the family-wise error rate was controlled at $\alpha = 0.05$ with the Holm (1979) step-down procedure, Eq. (10), where $p_{(1)} \le \ldots \le p_{(m)}$ are the ordered p-values and m = 15:

$$p_{(i)}^{\text{Holm}} = \max_{j \le i} \min\{1,\ (m - j + 1)\, p_{(j)}\} \tag{10}$$

Calibration. Unbiasedness was assessed with the Mincer–Zarnowitz (MZ) regression (Mincer & Zarnowitz, 1969), Eq. (11), testing the joint hypothesis $a = 0$ and $b = 1$ with a Wald statistic based on the Newey–West covariance matrix with 5 lags (Newey & West, 1987). A slope $b > 1$ indicates forecasts that vary too little. Remaining serial dependence in the forecast errors was tested with the Ljung–Box statistic, Eq. (2), with m = 10.

$$y_t = a + b\,\hat{y}_t + v_t, \qquad H_0: a = 0,\ b = 1 \tag{11}$$

3.6. *PV potential estimation*

Because radiation in kWh/m²/day is equivalent to peak sun hours, the daily PV specific yield was estimated as in Eq. (12), where H is the daily horizontal radiation and PR is the

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

performance ratio, assumed to be 0.80. This estimate ignores module tilt and shading and is therefore indicative for comparisons between cities.

$$Y_f = H \times \mathrm{PR} \tag{12}$$

4. Results and Discussion

4.1. *Radiation characteristics and time-series properties*

Kupang received the highest mean radiation, 5.93 kWh/m²/day, and was the most stable, with a coefficient of variation (CV) of 19.1%. Jayapura received the lowest, 4.91 kWh/m²/day, and was the most variable, with a CV of 27.2% (Table 1). The ranking follows cloud cover: the ratio of all-sky to clear-sky radiation over 2014–2023 was highest in Kupang (0.86), the driest city, with an annual rainfall of 1,398 mm, and lowest in Jayapura (0.71) and Ambon (0.72), the two wettest cities. Days with radiation below 3 kWh/m² numbered only 88 (2.4%) in Kupang but 392 (10.7%) in Jayapura.

The monthly means show three seasonal patterns. Makassar and Kupang peak at the end of the dry season in September–October (6.52 and 7.07 kWh/m²/day). Radiation in Makassar is lowest during the December–January rainy season (4.74 and 4.58), whereas Kupang reaches its minimum in June (4.95), a dip that coincides with its lowest clear-sky radiation and is therefore more likely due to solar geometry at 10°S than to cloud cover. Ambon shows the opposite, local pattern, with a minimum of 3.70 kWh/m²/day in July during the June–July rainfall peak (11.7 and 11.2 mm/day), consistent with the Moluccan rainfall region (Aldrian & Susanto, 2003). Manado and Jayapura are flatter, with monthly amplitudes of 1.27 and 0.87 kWh/m²/day, compared with 1.94–2.12 in the other three cities. The lowest annual means in Makassar, Kupang, and Manado occurred in 2022, during the 2020–2022 La Niña, which tends to increase rainfall in Indonesia (Hendon, 2003).

Table 1 also summarizes the time-series diagnostics of the training data. Two Fourier harmonics were selected by BIC in every city, but the share of daily variance explained by the seasonal cycle ranged from 32.6% in Kupang to only 2.1% in Jayapura. After removal of this cycle, the ADF statistics (−15.66 to −17.71) were far below the 1% critical value of −3.43, so a unit root was rejected in every city and the anomalies were modeled without differencing. The KPSS statistics nevertheless exceeded the 5% critical value in Makassar, Ambon, and Manado and the 10% value in Kupang and Jayapura. The two tests together indicate an anomaly series without a stochastic trend but with slowly varying levels, such as the wetter 2020–2022 period, which is a reason to treat the 2023 test year as a genuinely out-of-sample period.

Table 1: Study cities, mean radiation and PV specific yield (2014–2023), and time-series diagnostics of the daily radiation anomalies (training data 2014–2021, n = 2,922).

City	Coordinates	H	CV (%)	Y_f	R^2_{clim}	ACF(1)	ADF τ (lags)	KPSS η_μ
Kupang	10.17°S, 123.61°E	5.93	19.1	4.75	0.326	0.555	−16.92 (3)	0.376[a]
Makassar	5.15°S, 119.43°E	5.40	23.2	4.32	0.289	0.437	−15.66 (6)	0.602[b]
Ambon	3.70°S, 128.18°E	4.99	25.5	3.99	0.253	0.372	−17.71 (4)	0.972[c]
Manado	1.47°N, 124.84°E	5.42	21.3	4.34	0.124	0.353	−17.38 (4)	0.633[b]
Jayapura	2.53°S, 140.72°E	4.91	27.2	3.93	0.021	0.093	−15.95 (7)	0.354[a]

Note: H = mean daily radiation (kWh/m²/day); CV = coefficient of variation; Y_f = mean PV specific yield with PR = 0.80 (kWh/kWp/day); R^2_{clim} = share of daily variance explained by the Fourier climatology (two harmonics, selected by BIC); ACF(1) = lag-1 autocorrelation of the anomalies. ADF 1% critical value −3.43 (MacKinnon, 1996). KPSS critical values 0.347 (10%), 0.463 (5%), and 0.739 (1%); [a] $p < 0.10$, [b] $p < 0.05$, [c] $p < 0.01$. Ljung–Box $Q(10)$ of the anomalies ranged from 109 (Jayapura) to 2,054 (Kupang), all $p < 0.001$.

Serial dependence differed sharply between cities. The lag-1 autocorrelation of the anomalies was 0.555 in Kupang, 0.437 in Makassar, 0.372 in Ambon, 0.353 in Manado, and only 0.093 in Jayapura, and the Ljung–Box statistic fell from 2,054 to 109 in the same order. Because one-day-ahead forecasts can exploit only this dependence and the seasonal cycle, Jayapura was expected to be the most difficult city to forecast regardless of the model.

4.2. *Forecast accuracy*

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 2: Forecast accuracy on the 2023 test data (n = 365 days; RMSE, MAE, and ME in kWh/m²/day).

City	Model	RMSE, 5 seeds (mean ± SD)	RMSE [95% CI]	MAE	R^2	SS	ME
Kupang	Persistence	–	0.789 [0.653, 0.917]	0.497	0.530	0.000	0.005
	Climatology	–	0.932 [0.748, 1.109]	0.618	0.344	−0.182	−0.017
	DHR-ARMA	–	0.722 [0.596, 0.843]	0.465	0.606	0.085	0.000
	ARX	–	**0.717 [0.597, 0.831]**	0.479	0.612	0.091	0.057
	LSTM	0.730 ± 0.004	0.731 [0.610, 0.847]	0.483	0.596	0.074	0.028
	GRU	0.734 ± 0.010	0.751 [0.642, 0.857]	0.542	0.574	0.048	0.151
	BiLSTM	0.730 ± 0.006	0.739 [0.613, 0.859]	0.481	0.587	0.063	0.027
Makassar	Persistence	–	0.976 [0.815, 1.129]	0.666	0.392	0.000	−0.003
	Climatology	–	0.962 [0.837, 1.094]	0.750	0.408	0.014	0.127
	DHR-ARMA	–	0.824 [0.704, 0.940]	0.615	0.567	0.156	0.061
	ARX	–	0.818 [0.704, 0.932]	0.625	0.573	0.162	0.094
	LSTM	0.810 ± 0.004	0.807 [0.692, 0.926]	0.614	0.584	0.173	0.106
	GRU	0.800 ± 0.004	**0.802 [0.697, 0.908]**	0.627	0.589	0.178	0.141
	BiLSTM	0.809 ± 0.007	0.806 [0.702, 0.911]	0.631	0.585	0.174	0.166
Ambon	Persistence	–	1.172 [1.061, 1.282]	0.887	0.174	0.000	0.008
	Climatology	–	1.045 [0.960, 1.132]	0.857	0.342	0.108	0.035
	DHR-ARMA	–	0.969 [0.887, 1.052]	0.767	0.434	0.173	0.023
	ARX	–	0.972 [0.891, 1.052]	0.788	0.431	0.170	0.062
	LSTM	0.951 ± 0.004	0.956 [0.874, 1.036]	0.769	0.450	0.184	0.090
	GRU	0.952 ± 0.004	0.958 [0.874, 1.039]	0.764	0.448	0.183	0.072
	BiLSTM	0.950 ± 0.002	**0.951 [0.870, 1.030]**	0.763	0.456	0.188	0.129
Manado	Persistence	–	1.224 [1.045, 1.391]	0.822	0.021	0.000	0.004
	Climatology	–	1.138 [1.005, 1.274]	0.869	0.153	0.070	0.056
	DHR-ARMA	–	1.010 [0.877, 1.136]	0.736	0.333	0.175	0.030
	ARX	–	1.007 [0.876, 1.132]	0.750	0.338	0.177	0.084
	LSTM	0.984 ± 0.007	**0.979 [0.856, 1.098]**	0.735	0.373	0.200	0.100
	GRU	0.985 ± 0.002	0.986 [0.865, 1.104]	0.745	0.364	0.194	0.118
	BiLSTM	0.983 ± 0.006	0.982 [0.862, 1.097]	0.745	0.370	0.198	0.125
Jayapura	Persistence	–	1.802 [1.654, 1.947]	1.408	−0.816	0.000	0.004
	Climatology	–	1.327 [1.237, 1.416]	1.079	0.015	0.263	−0.066
	DHR-ARMA	–	1.321 [1.231, 1.412]	1.075	0.024	0.267	−0.042
	ARX	–	1.315 [1.224, 1.404]	1.061	0.033	0.270	−0.047
	LSTM	1.303 ± 0.004	1.298 [1.210, 1.385]	1.060	0.058	0.280	−0.009
	GRU	1.304 ± 0.002	1.303 [1.214, 1.390]	1.055	0.051	0.277	−0.026
	BiLSTM	1.294 ± 0.003	**1.294 [1.203, 1.383]**	1.046	0.064	0.282	−0.036

Note: For the networks, the first RMSE column summarizes the five seeds, and all other columns refer to the seed with the lowest validation loss. Intervals: circular block bootstrap, block length 7 days, 10,000 replicates. SS = skill score relative to persistence; ME = mean error (actual − forecast). Bold marks the lowest RMSE in each city.

Table 2 reports the accuracy of all seven forecasting methods on the 2023 test year, and Figure 1 displays the RMSE with its 95% block-bootstrap interval. The three networks and the two linear models formed a tight cluster in every city: their RMSEs spanned 0.717–0.751 in Kupang, 0.802–0.824 in Makassar, 0.951–0.972 in Ambon, 0.979–1.010 in Manado, and 1.294–1.321 kWh/m²/day in Jayapura. Within each cluster, the spread was at most 0.034 kWh/m²/day, whereas the bootstrap intervals were 0.16–0.26 kWh/m²/day wide.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

ARX gave the lowest RMSE in Kupang; elsewhere, the best network was 1.6–2.7% more accurate than ARX.

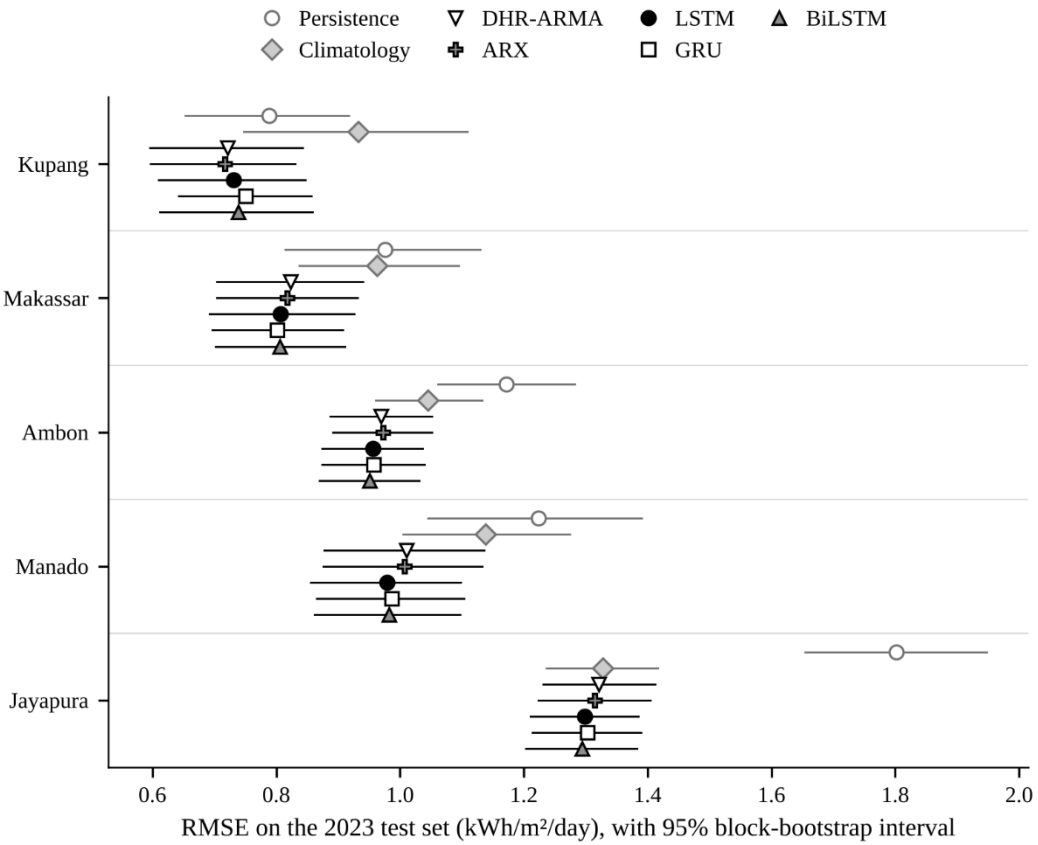


Figure 1: RMSE of the seven forecasting methods on the 2023 test data, with 95% circular block-bootstrap intervals.

Between cities, by contrast, the differences were large. The best RMSE rose from 0.717 kWh/m²/day in Kupang to 1.294 in Jayapura, and the best R^2 fell from 0.612 to 0.064, in the same order as the anomaly autocorrelation in Table 1. The two extremes illustrate this: in Kupang, the BiLSTM forecasts follow the day-to-day fluctuations (correlation 0.77), whereas in Jayapura they stay close to the mean, with a standard deviation of 0.39 kWh/m²/day against 1.34 for the observations and a correlation of 0.26.

Relative to persistence, the skill scores of the networks and linear models were 0.05–0.09 in Kupang, where persistence already achieved $R^2 = 0.530$, and 0.16–0.28 in the other cities, rising to 0.27–0.28 in Jayapura, where persistence was very poor ($R^2 = -0.816$). Relative to monthly climatology, the best method reduced the RMSE by 23.1% in Kupang, 16.7% in Makassar, 14.0% in Manado, and 9.0% in Ambon, but by only 2.5% in Jayapura (1.294 vs. 1.327). In Jayapura, DHR-ARMA (1.321) was practically identical to climatology, consistent with an anomaly autocorrelation of 0.093.

Errors also depended on season and cloud conditions. For the best network by seed-averaged RMSE, the RMSE in Kupang was 1.10 kWh/m²/day in December–February but 0.22 in September–November; in Makassar, the corresponding values were 1.08 and 0.47. The largest 10% of errors in each city occurred on cloudier days, with a mean actual radiation of 2.58–4.50 kWh/m²/day, and on these days radiation was overestimated in 62–92% of cases. Model complexity, in contrast to accuracy, differed by orders of magnitude: ARX has 87 parameters and DHR-ARMA 7–10, compared with 14,273 for GRU, 18,753 for LSTM, and 37,505 for BiLSTM, whose mean training times were 16.5, 33.5, and 52.9 s per run on a single CPU core.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

4.3. *Sources of variation in accuracy*

The two-way ANOVA in Table 3 quantifies the visual impression of Figure 1. City accounted for 99.92% of the total sum of squares of the 75 network RMSEs ($F(4, 60) = 27{,}317.9$, $p < 0.001$). The main effect of architecture was not significant ($F(2, 60) = 1.50$, $p = 0.231$) and accounted for 0.003% of the variation. The city × architecture interaction was significant ($F(8, 60) = 2.81$, $p = 0.010$) but accounted for only 0.021%. Seed-to-seed variability, with standard deviations of 0.002–0.010 kWh/m²/day within cells, accounted for the remaining 0.055%.

Table 3: Two-way ANOVA of test RMSE for the recurrent networks (5 cities × 3 architectures × 5 seeds).

Source	SS	df	MS	*F*	*p*	Share of total SS (%)	Partial η^2
City	2.882199	4	0.7205498	27,317.9	<0.001	99.922	0.999
Model	0.000079	2	0.0000396	1.50	0.231	0.003	0.048
City × Model	0.000594	8	0.0000742	2.81	0.010	0.021	0.273
Residual (seed)	0.001583	60	0.0000264			0.055	
Total	2.884455	74				100	

Note: Brown–Forsythe test of equal variances: $F = 1.17$, $p = 0.324$. Shapiro–Wilk test of residuals: $W = 0.965$, $p = 0.037$. ANOVA on log RMSE: architecture $p = 0.506$; interaction $p = 0.035$.

The assumption of equal variances was satisfied (Brown–Forsythe $p = 0.324$). The residuals departed mildly from normality (Shapiro–Wilk $p = 0.037$), but the ANOVA on log RMSE led to the same conclusions (architecture $p = 0.506$; interaction $p = 0.035$). One-way ANOVAs within each city showed that the interaction came from Jayapura ($F(2, 12) = 19.52$, $p < 0.001$) and Makassar ($F(2, 12) = 5.17$, $p = 0.024$), where the seed-averaged RMSEs of the three architectures differed by 0.010 and 0.009 kWh/m²/day; in the other three cities, $p \geq 0.62$. These differences are detectable only because seed variability is very small; they amount to about 4–6% of the width of the bootstrap intervals in Table 2 and therefore have little practical meaning.

4.4. *Tests of equal predictive accuracy*

No pair of networks differed significantly in any city, even before adjustment: the largest DM values were −1.82 for LSTM versus GRU in Kupang and 1.81 for GRU versus BiLSTM in Jayapura (unadjusted $p = 0.070$ and 0.071), and all 15 Holm-adjusted p-values in family A equalled 1. ARX and DHR-ARMA were likewise equally accurate in every city (all unadjusted $p \geq 0.69$). Table 4 compares each network with the benchmarks. Every network significantly outperformed persistence in Makassar, Ambon, Manado, and Jayapura (Holm-adjusted $p \leq 0.001$), and so did ARX and DHR-ARMA (family E, Holm-adjusted $p \leq 0.002$). In Kupang, no model outperformed persistence significantly (Holm-adjusted $p \geq 0.40$).

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 4: Diebold–Mariano tests of the networks against persistence, ARX, and DHR-ARMA: DM (unadjusted p; Holm-adjusted p).

City	Model	vs. persistence	vs. ARX	vs. DHR-ARMA
Kupang	LSTM	−1.42 (0.156; 0.468)	1.22 (0.222; 1.000)	0.62 (0.536; 1.000)
	GRU	−1.00 (0.317; 0.468)	2.66 (0.008; 0.115)	1.86 (0.063; 0.820)
	BiLSTM	−1.31 (0.192; 0.468)	1.98 (0.049; 0.536)	1.14 (0.254; 1.000)
Makassar	LSTM	−3.71 (<0.001; <0.001)	−0.81 (0.417; 1.000)	−0.94 (0.348; 1.000)
	GRU	−4.17 (<0.001; <0.001)	−1.63 (0.103; 0.930)	−1.60 (0.111; 1.000)
	BiLSTM	−4.21 (<0.001; <0.001)	−1.07 (0.286; 1.000)	−1.38 (0.170; 1.000)
Ambon	LSTM	−5.10 (<0.001; <0.001)	−1.02 (0.307; 1.000)	−0.64 (0.526; 1.000)
	GRU	−4.91 (<0.001; <0.001)	−0.86 (0.389; 1.000)	−0.53 (0.595; 1.000)
	BiLSTM	−5.20 (<0.001; <0.001)	−1.26 (0.208; 1.000)	−0.84 (0.400; 1.000)
Manado	LSTM	−4.25 (<0.001; <0.001)	−2.75 (0.006; 0.093)	−1.93 (0.054; 0.808)
	GRU	−4.15 (<0.001; <0.001)	−2.25 (0.025; 0.303)	−1.49 (0.136; 1.000)
	BiLSTM	−4.23 (<0.001; <0.001)	−2.51 (0.013; 0.165)	−1.78 (0.076; 0.912)
Jayapura	LSTM	−7.52 (<0.001; <0.001)	−1.37 (0.170; 1.000)	−1.60 (0.112; 1.000)
	GRU	−7.42 (<0.001; <0.001)	−1.16 (0.247; 1.000)	−1.24 (0.215; 1.000)
	BiLSTM	−7.55 (<0.001; <0.001)	−1.88 (0.061; 0.614)	−1.93 (0.054; 0.808)

Note: A negative DM means that the first model (network) is more accurate. Holm adjustment within each family of 15 tests (five cities × three comparisons).

Against the linear models, five of the 30 unadjusted tests were significant at 5%, in two opposite directions. In Manado, all three networks were more accurate than ARX (p = 0.006–0.025), with bootstrap differences in RMSE of −0.020 to −0.028 kWh/m²/day whose intervals excluded zero. In Kupang, ARX was more accurate than GRU (p = 0.008; difference 0.034, 95% CI 0.011–0.059) and BiLSTM (p = 0.049; 0.022, CI 0.001–0.043). None of these differences survived the Holm correction (smallest adjusted p = 0.093 against ARX and 0.808 against DHR-ARMA). With 15 tests per family and no true differences, about 0.75 unadjusted rejections per family would be expected by chance, so the scattered and sign-inconsistent rejections are better read as sampling noise than as evidence of superiority.

4.5. *Forecast calibration*

Table 5 reports the calibration diagnostics. In Makassar, Ambon, and Manado and for GRU in Kupang, the networks underestimated radiation on average by 0.07–0.17 kWh/m²/day, about 1.4–3.0% of the 2023 mean. The MZ test rejected unbiasedness for 9 of the 15 networks: all three in Makassar and Manado, LSTM and GRU in Kupang, and BiLSTM in Ambon. In eight of these nine cases, the slope was 1.06–1.20 and the intercept negative, which indicates forecasts that are shrunk toward the mean, as expected from networks trained to minimize squared error; BiLSTM in Ambon was rejected because of its bias alone. The linear models were better calibrated: unbiasedness was rejected for ARX in Manado and Jayapura and for DHR-ARMA only in Makassar, and the mean errors of DHR-ARMA did not exceed 0.061 kWh/m²/day in absolute value.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 5: Forecast-error diagnostics on the 2023 test data.

City	Model	ME	MZ intercept (SE)	MZ slope (SE)	Wald $\chi^2(2)$ [p]	Ljung–Box $Q(10)$ [p]
Kupang	LSTM	0.028	−0.591 (0.305)	1.105 (0.047)	12.97 [0.002]	68.78 [<0.001]
	GRU	0.151	−0.559 (0.312)	1.123 (0.049)	50.59 [<0.001]	77.08 [<0.001]
	BiLSTM	0.027	−0.159 (0.310)	1.031 (0.048)	2.23 [0.327]	69.65 [<0.001]
	ARX	0.057	0.011 (0.256)	1.008 (0.039)	4.01 [0.135]	55.16 [<0.001]
	DHR-ARMA	0.000	−0.248 (0.263)	1.042 (0.040)	1.49 [0.475]	46.73 [<0.001]
Makassar	LSTM	0.106	−0.324 (0.326)	1.079 (0.054)	21.46 [<0.001]	10.44 [0.403]
	GRU	0.141	−0.301 (0.265)	1.082 (0.045)	28.94 [<0.001]	6.82 [0.742]
	BiLSTM	0.166	−0.185 (0.251)	1.065 (0.042)	35.59 [<0.001]	4.30 [0.933]
	ARX	0.094	0.175 (0.256)	0.985 (0.042)	4.79 [0.091]	11.83 [0.297]
	DHR-ARMA	0.061	−0.400 (0.287)	1.084 (0.048)	10.62 [0.005]	5.10 [0.884]
Ambon	LSTM	0.090	−0.148 (0.316)	1.048 (0.062)	3.97 [0.137]	13.94 [0.176]
	GRU	0.072	−0.102 (0.304)	1.035 (0.060)	2.50 [0.287]	14.08 [0.169]
	BiLSTM	0.129	0.080 (0.290)	1.010 (0.057)	6.70 [0.035]	12.68 [0.242]
	ARX	0.062	−0.130 (0.340)	1.039 (0.066)	2.00 [0.368]	14.00 [0.173]
	DHR-ARMA	0.023	−0.214 (0.335)	1.047 (0.064)	0.98 [0.613]	13.46 [0.199]
Manado	LSTM	0.100	−0.983 (0.419)	1.200 (0.074)	16.43 [<0.001]	20.40 [0.026]
	GRU	0.118	−0.864 (0.405)	1.182 (0.072)	17.18 [<0.001]	20.96 [0.021]
	BiLSTM	0.125	−0.865 (0.414)	1.184 (0.073)	19.10 [<0.001]	19.28 [0.037]
	ARX	0.084	−0.534 (0.427)	1.114 (0.076)	6.07 [0.048]	20.27 [0.027]
	DHR-ARMA	0.030	−0.641 (0.411)	1.122 (0.073)	3.79 [0.151]	24.18 [0.007]
Jayapura	LSTM	−0.009	0.662 (0.728)	0.861 (0.150)	0.86 [0.651]	6.84 [0.740]
	GRU	−0.026	0.771 (0.764)	0.836 (0.156)	1.28 [0.528]	6.33 [0.787]
	BiLSTM	−0.036	0.565 (0.705)	0.877 (0.144)	1.05 [0.591]	6.48 [0.773]
	ARX	−0.047	1.679 (0.627)	0.646 (0.128)	8.11 [0.017]	6.30 [0.789]
	DHR-ARMA	−0.042	0.830 (1.289)	0.821 (0.264)	0.83 [0.660]	7.97 [0.632]

Note: ME = mean error (actual − forecast). MZ = Mincer–Zarnowitz regression; Wald test of $a = 0$, $b = 1$ with Newey–West covariance (5 lags). Ljung–Box test on forecast errors with 10 lags.

The Ljung–Box test detected remaining autocorrelation in the forecast errors of all five models in Kupang ($Q(10) = 46.7–77.1$, $p < 0.001$) and Manado ($p = 0.007–0.037$), but not in the other three cities ($p \geq 0.17$). Because all models shared this pattern, the residual dependence reflects information missing from the inputs rather than a deficiency of one architecture. In Kupang, the city with the most persistent radiation, the networks also used this persistence less effectively than the linear models, consistent with ARX and DHR-ARMA ranking first and second there.

4.6. *Implications for PV potential*

Converting radiation into PV specific yield with PR = 0.80 gives mean values of 4.75 kWh/kWp/day in Kupang, 4.34 in Manado, 4.32 in Makassar, 3.99 in Ambon, and 3.93 in Jayapura (Table 1). The specific yield in Kupang is thus about 21% higher than in Jayapura, a range slightly above the 3.4–4.1 kWh/kWp/day reported by Silalahi et al. (2021) for 34 Indonesian cities, as expected from the assumed PR and the tendency of satellite data to overestimate radiation (Harsarapama et al., 2020). A 1 MWp plant would produce about 1,734 MWh per year in Kupang but about 1,434 MWh in Jayapura, and in Ambon the July output is only about 64% of the November output.

The calibration results translate into energy: for 2023, the networks underestimated annual energy by 0.5–3.0% in four cities and overestimated it by 0.2–0.7% in Jayapura. Expressed in energy units, the daily error of the best network (seed-averaged RMSE × PR) is about 0.58 kWh/kWp in Kupang, 0.64 in Makassar, 0.76 in Ambon, 0.79 in Manado, and 1.03 in Jayapura, or 12–27% of the mean daily production; these values can serve as an initial reference for sizing reserves or storage.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

4.7. Discussion

The central result is that the location, not the model, determines forecast accuracy. City explained 99.9% of the variation in network RMSE, whereas architecture had no significant main effect, and the significant interaction amounted to differences of about 0.01 kWh/m²/day. This agrees with Greff et al. (2017), who found that none of eight LSTM variants significantly outperformed the standard LSTM, and with earlier reports of close performance among LSTM, GRU, and BiLSTM (Zameer et al., 2023). The city ranking followed the autocorrelation of the anomalies, a property of the climate that no architecture can change.

The comparison with linear models sharpens this conclusion. ARX uses the same 14-day inputs as the networks but no nonlinearity, and its RMSE was within 3% of the best network in every city; after control of the family-wise error rate, none of the 30 differences between networks and linear models was significant. At a one-day horizon with daily inputs, the additional capacity of recurrent networks therefore adds little that a well-specified linear time-series model cannot capture, in line with the evidence of Makridakis et al. (2018). The unadjusted advantage of the networks in Manado suggests some nonlinear structure in a wet near-equatorial climate, whereas the advantage of ARX in Kupang suggests that early-stopped networks may underuse strong persistence; both effects are small relative to the sampling uncertainty of one test year.

Two sources of uncertainty must be distinguished. Seed variability (standard deviations of 0.002–0.010 kWh/m²/day) is the uncertainty captured when results are reported as mean ± standard deviation over runs, and it is an order of magnitude smaller than the sampling variability of the test days, which the bootstrap intervals (0.16–0.26 kWh/m²/day wide) and the DM test capture. A comparison based only on seed-averaged RMSE, as is common in deep learning studies, can therefore show statistically detectable differences, such as those in Jayapura and Makassar here, that do not hold up against day-to-day variation in the test period. Without multiplicity control, the 75 DM tests in this study would also have produced five apparent wins or losses between networks and linear models.

The calibration results add a dimension that accuracy measures conceal. The networks were not only no more accurate than the linear models but also more often biased, with forecasts shrunk toward the mean (MZ slopes of 1.06–1.20 in eight of the nine rejected cases). For PV planning, this produces a systematic underestimation of energy of up to 3% per year. A linear recalibration of the network forecasts using the MZ regression estimated on validation data is a simple remedy that should be evaluated in future work.

The main source of error, sudden overcast days, cannot be inferred from the radiation and surface weather of preceding days, which also explains why reading the window backward gives BiLSTM no advantage. Overcast days largely coincide with rainy days: in the NASA POWER data, 82–100% of the days in the lowest radiation decile of each city had at least 1 mm of rain. Further gains are therefore more likely from new information, such as numerical weather predictions (Qing & Niu, 2018), satellite cloud imagery, data from surrounding locations (Brahma & Wadhvani, 2020; Kumari & Toshniwal, 2021), or next-day rain forecasts from rainfall classifiers already developed for the region (Hafid et al., 2025; Rais, Hafid, et al., 2026), than from more complex networks. For operational use, DHR-ARMA and ARX offer equal accuracy with interpretable parameters and better calibration; among the networks, GRU is the most economical. In Jayapura, none of the models adds much beyond climatology, so PV planning there should use wide uncertainty ranges.

This study has several limitations. First, NASA POWER is a coarse satellite and reanalysis product, which tends to overestimate radiation in Indonesia by 3–7% (Harsarapama et al., 2020) and should be validated against BMKG stations or local pyranometers. Second, a single test year of 365 days limits statistical power: differences of about 2% in RMSE cannot be detected reliably, and the Holm procedure is conservative. Third, the network

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

hyperparameters were fixed across architectures, the DM tests used one seed per network, and the ARMA parameters were estimated by conditional least squares in two stages. Fourth, the KPSS results indicate level shifts that a fixed climatology does not capture, and other horizons or ENSO phases may change the ranking. Finally, the energy estimates use horizontal radiation and a fixed PR.

5. Conclusion

This study compared three recurrent neural networks (LSTM, GRU, and BiLSTM) with two linear time-series models (DHR-ARMA and ARX) for one-day-ahead forecasting of daily solar radiation in Makassar, Kupang, Ambon, Manado, and Jayapura, using NASA POWER data for 2014–2023 and a statistical evaluation framework.

Accuracy was determined by location rather than by model. In a two-way ANOVA of 75 network runs, city explained 99.9% of the variation in RMSE, and architecture had no significant effect ($p = 0.231$). The best RMSE ranged from 0.717 kWh/m²/day in Kupang to 1.294 in Jayapura, following the day-to-day autocorrelation of radiation anomalies (0.555 to 0.093).

No network was significantly more accurate than another network or than the linear models after Holm correction (45 tests), and the linear models were within 3.2% of the best network and best in Kupang. All models significantly outperformed persistence in four cities but not in Kupang. The networks were more often biased than the linear models, underestimating radiation by up to 0.17 kWh/m²/day and annual PV energy by up to 3%. The specific yield ranged from 3.93 kWh/kWp/day in Jayapura to 4.75 in Kupang.

Methodologically, the results show that deep learning forecasts should be evaluated against linear models with the same inputs, with multiplicity-adjusted tests that account for the sampling variability of the test period rather than only the variability across training runs, and with calibration diagnostics. Future work should add input information, such as weather predictions or satellite cloud data, recalibrate network forecasts, and validate the results against ground measurements.

References

Ahmar, A. S., Rais, Z., & Tunnas, F. (2025). Comparative analysis of support vector regression (SVR) and SutteARIMA in predicting coal prices in Indonesia. *Quality & Quantity*, *59*(5), 4185–4200. https://doi.org/10.1007/s11135-025-02162-2

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Aldrian, E., & Susanto, R. D. (2003). Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. *International Journal of Climatology*, *23*, 1435–1452. https://doi.org/10.1002/joc.950

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., & Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, *136*, 78–111. https://doi.org/10.1016/j.solener.2016.06.069

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Brahma, B., & Wadhvani, R. (2020). Solar irradiance forecasting based on deep learning methodologies and multi-site data. *Symmetry*, *12*(11), 1830. https://doi.org/10.3390/sym12111830

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427–431. https://doi.org/10.1080/01621459.1979.10482531

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Ghimire, S., Deo, R. C., Raj, N., & Mi, J. (2019). Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. *Applied Energy*, *253*, 113541. https://doi.org/10.1016/j.apenergy.2019.113541

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5–6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Hafid, H., Rais, Z., & Rezky, A. R. R. T. (2025). Klasifikasi curah hujan di Kota Makassar menggunakan gradient boosting machine (GBM) [Rainfall classification in Makassar City using a gradient boosting machine (GBM)]. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *7*(2), 133–142. https://doi.org/10.35580/variansiunm386

Harsarapama, A. P., Aryani, D. R., & Rachmansyah, D. (2020). Open-source satellite-derived solar resource databases comparison and validation for Indonesia. *Journal of Renewable Energy*, *2020*. https://doi.org/10.1155/2020/2134271

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291.

Hendon, H. H. (2003). Indonesian rainfall variability: Impacts of ENSO and local air–sea interaction. *Journal of Climate*, *16*, 1775–1790.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. arXiv:1412.6980.

Kumari, P., & Toshniwal, D. (2021). Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. *Applied Energy*, *295*, 117061. https://doi.org/10.1016/j.apenergy.2021.117061

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241. https://doi.org/10.1214/aos/1176347265

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Langer, J., Quist, J., & Blok, K. (2021). Review of renewable energy potentials in Indonesia and their contribution to a 100% renewable electricity system. *Energies*, *14*(21), 7033. https://doi.org/10.3390/en14217033

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297

MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*, *11*(6), 601–618. https://doi.org/10.1002/(SICI)1099-1255(199611)11:6%3C601::AID-JAE417%3E3.0.CO;2-T

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE*, *13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Marquez, R., & Coimbra, C. F. M. (2013). Proposed metric for evaluation of solar forecasting models. *Journal of Solar Energy Engineering*, *135*(1), 011016. https://doi.org/10.1115/1.4007496

Mincer, J. A., & Zarnowitz, V. (1969). The evaluation of economic forecasts. In J. A. Mincer (Ed.), *Economic forecasts and expectations: Analysis of forecasting behavior and performance* (pp. 3–46). National Bureau of Economic Research.

Montgomery, D. C. (2017). *Design and analysis of experiments* (9th ed.). Wiley.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708. https://doi.org/10.2307/1913610

Politis, D. N., & Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In R. LePage & L. Billard (Eds.), *Exploring the limits of bootstrap* (pp. 263–270). Wiley.

Pradipta, M. I., & Haviana, S. F. C. (2025). Prediksi penghematan biaya listrik berdasarkan global horizontal irradiance (GHI) menggunakan model long short term memory network (LSTM) [Prediction of electricity cost savings based on global horizontal irradiance (GHI) using a long short-term memory network (LSTM) model]. *TRANSISTOR Elektro dan Informatika*, *7*(2), 59–69. https://jurnal.unissula.ac.id/index.php/EI/article/view/44570

Qing, X., & Niu, Y. (2018). Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy*, *148*, 461–468. https://doi.org/10.1016/j.energy.2018.01.177

Rahmat, R. W., Annas, S., & Rais, Z. (2023). Analisis support vector regression (SVR) untuk meramalkan indeks kualitas udara di Kota Makassar [Support vector regression (SVR) analysis for forecasting the air quality index in Makassar City]. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *5*(3), 104–117. https://jurnalvariansi.unm.ac.id/index.php/variansi/article/view/107

Rais, Z., Hafid, H., & Bunga, Y. R. (2026). Rainfall classification using output statistics models based on classification and regression trees with principal component analysis preprocessing. *JINAV: Journal of Information and Visualization*, *7*(1), 69–82. https://doi.org/10.35877/454RI.jinav4788

Rais, Z., Sumarno, A. S., & Utomo, A. T. (2026). Forecasting IDR exchange rate to USD using hybrid ARIMA–LSTM. *ARRUS Journal of Engineering and Technology*, *6*(1), 11–17. https://journal.arrus.id/index.php/jetech/article/view/4899

Ruliana, R., Maru, R., Rais, Z., & Ahmar, A. S. (2025). Integration of SEM and nonparametric spline in spatial data modeling and visualization for analysis of CO2 reduction based on green space in Makassar City. *JOIV: International Journal on Informatics Visualization*, *9*(3), 1319–1329. https://doi.org/10.62527/joiv.9.3.3807

Rutan, D. A., Kato, S., Doelling, D. R., Rose, F. G., Nguyen, L. T., Caldwell, T. E., & Loeb, N. G. (2015). CERES synoptic product: Methodology and validation of surface radiant flux. *Journal of Atmospheric and Oceanic Technology*, *32*(6), 1121–1143. https://doi.org/10.1175/JTECH-D-14-00165.1

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5607

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Silalahi, D. F., Blakers, A., Stocks, M., Lu, B., Cheng, C., & Hayes, L. (2021). Indonesia's vast solar energy potential. *Energies*, *14*(17), 5424. https://doi.org/10.3390/en14175424

Sparks, A. H. (2018). nasapower: A NASA POWER global meteorology, surface solar energy and climatology data client for R. *Journal of Open Source Software*, *3*(30), 1035. https://doi.org/10.21105/joss.01035

Veldhuis, A. J., & Reinders, A. H. M. E. (2013). Reviewing the potential and cost-effectiveness of grid-connected solar PV in Indonesia on a provincial level. *Renewable and Sustainable Energy Reviews*, *27*, 315–324. https://doi.org/10.1016/j.rser.2013.06.010

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582.

Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2019). A review of deep learning for renewable energy forecasting. *Energy Conversion and Management*, *198*, 111799. https://doi.org/10.1016/j.enconman.2019.111799

Wojtkiewicz, J., Hosseini, M., Gottumukkala, R., & Chambers, T. L. (2019). Hour-ahead solar irradiance forecasting using multivariate gated recurrent units. *Energies*, *12*(21), 4055. https://doi.org/10.3390/en12214055

Yang, D., Kleissl, J., Gueymard, C. A., Pedro, H. T. C., & Coimbra, C. F. M. (2018). History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, *168*, 60–101. https://doi.org/10.1016/j.solener.2017.11.023

Zameer, A., Jaffar, F., Shahid, F., Muneeb, M., Khan, R., & Nasir, R. (2023). Short-term solar energy forecasting: Integrated computational intelligence of LSTMs and GRU. *PLoS ONE*, *18*(10), e0285410. https://doi.org/10.1371/journal.pone.0285410

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.