



Available online at <http://scik.org>

Commun. Math. Biol. Neurosci. 2026, 2026:96

<https://doi.org/10.28919/cmbn/10088>

ISSN: 2052-2541

INTEGRATING TIME-LAG EFFECTS IN SEMIPARAMETRIC TIME SERIES REGRESSION MODELING OF RELATIVE HUMIDITY USING FOURIER SERIES AND LOCAL POLYNOMIAL: A STUDY BASED ON MERRA 2 REANALYSIS AND SATELLITE DATA

ANDREA TRI RIAN DANI^{1,2}, MEYLITA SARI^{3*}, SYAIFUL ANAM³, HILMI AZIZ BUKHORI³,
AHMAD HAKIIM JAMALUDDIN⁴, YOSSY CANDRA⁵, NUR CHAMIDAH⁶, I NYOMAN BUDIANTARA⁷,
MEIRINDA FAUZIYAH¹, NAUFAL RAMADHAN AL-AKHWAL SIREGAR³

¹Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas
Mulawarman, Samarinda 75119, Indonesia

²Doctoral Study Program MIPA, Faculty of Science and Technology, Universitas Airlangga, Surabaya 60115,
Indonesia

³Department of Mathematics, Faculty of Science, Technology, and Mathematics, Universitas Brawijaya, Malang
65145, Indonesia

⁴Department of Mathematics & Statistics, Universiti Putra Malaysia, Selangor 43400, Malaysia

⁵Center for Religious Harmony, Secretariat General, Ministry of Religious Affairs, Jakarta 10710, Indonesia

⁶Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya 60115, Indonesia

⁷Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya
60111, Indonesia

Copyright © 2026 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits
unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: Nowadays, accurate modeling of relative humidity (RH) has become increasingly important for
hydrological, climatological, agricultural, and environmental applications. RH data typically exhibit strong seasonal

*Corresponding author

E-mail address: meylitasari@ub.ac.id

Received June 26, 2026

patterns, nonlinear behavior, and temporal dependence, making conventional parametric approaches less effective in capturing their underlying dynamics. This study aims to model RH using Fourier Series Nonparametric Regression (FSNR), Semiparametric Time Series Regression–Fourier Series (STSR–Fourier), and Semiparametric Time Series Regression–Local Polynomial (STSR–Local Polynomial) based on MERRA-2 reanalysis data collected from four observation stations in East Java, Indonesia, namely Juanda, Tanjung Perak, Karangploso, and Banyuwangi. The proposed STSR models incorporate a first-order time-lag component as the parametric part and either Fourier Series or Local Polynomial estimators as the nonparametric component. Model performance was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Normalized Mean Squared Error (NMSE), and Mean Absolute Percentage Error (MAPE). Three data-splitting schemes (95:5, 90:10, and 85:15) were examined, and the optimal scheme was selected using a majority-voting approach. The results revealed strong annual seasonality and significant temporal dependence in RH across all stations. The 95:5 data-splitting scheme achieved the highest number of wins and was therefore selected as the optimal scheme for model comparison. Comparative analysis showed that STSR–Fourier consistently produced lower testing errors than FSNR and STSR–Local Polynomial at most observation stations and achieved highly accurate predictions. The integration of time-lag effects and Fourier-based nonparametric estimation effectively captured both temporal dependence and seasonal fluctuations in RH data. These findings indicate that STSR–Fourier provides a flexible and reliable framework for modeling hydrometeorological time-series data characterized by nonlinear and periodic behavior.

Keywords: relative humidity; semiparametric time series regression; Fourier series; local polynomial; MERRA-2, time-lag effect.

2020 AMS Subject Classification: 62G08.

1. INTRODUCTION

Relative Humidity (RH) is one of the most important hydrometeorological variables influencing environmental processes, agricultural productivity, climatology, public health, and atmospheric systems [1]. Variations in RH affect cloud formation, the hydrological cycle, evapotranspiration processes, and weather dynamics. Consequently, accurate RH modeling is essential for supporting environmental monitoring, climate assessment, and data-driven decision-making in natural resource management [2], [3]. With the rapid advancement of remote sensing technology and climate reanalysis systems, hydrometeorological data with extensive spatial and temporal coverage have become increasingly available. Among these data sources, the Modern-

Era Retrospective Analysis for Research and Applications Version 2 (MERRA-2) has been widely utilized in atmospheric and climate studies because it provides continuous meteorological and hydrological observations with high temporal resolution [4], [5], [6], [7].

RH data typically exhibit complex time-series characteristics, including temporal dependence, seasonal fluctuations, and nonlinear patterns. Such characteristics often limit the effectiveness of conventional parametric approaches, which may fail to adequately represent the underlying data structure [8]. In recent years, the development of time-series modeling has accelerated, particularly with the increasing adoption of machine learning techniques. Despite their predictive capability, machine learning models are often criticized for their limited interpretability. In contrast, statistical approaches remain attractive because they provide interpretable model structures, inferential capabilities, and insights into the relationships among variables. One statistical framework that accommodates both linear and nonlinear relationships simultaneously is Semiparametric Time Series Regression (STSR) [9].

The STSR framework was first introduced by Truong [10], who proposed a time-series regression model that combines a nonparametric regression function with autoregressive error components. Subsequently, Gao developed semiparametric regression methods for nonlinear time-series modeling [11], which were later extended to semiparametric model selection through cross-validation procedures [12], nonparametric specification testing for nonstationary time-series data [13]. Furthermore, Chen et al. strengthened the theoretical foundation of STSR through the development of estimation procedures within the semiparametric time-series framework [14].

Several studies have further expanded the STSR framework through the development of different nonparametric estimators. Gao and Hawthorne introduced kernel estimators to address nonlinear relationships in time-series data [15]. Fibriyani et al. developed a Local Polynomial-based STSR model for inflation forecasting in Indonesia [16], while Fitriyah et al. proposed an STSR–Spline model for rice production modeling [17], [18]. More recently, Dani et al. introduced an STSR–Fourier model for land surface temperature modeling by integrating lag effects as the parametric component and Fourier Series functions as the nonparametric component [19]. These

studies demonstrate the flexibility of STSR in accommodating various forms of nonlinear relationships through different nonparametric estimation techniques.

Despite the growing development of STSR methodologies, several important research gaps remain. First, most previous studies have focused on economic and agricultural applications, whereas applications to hydrometeorological and atmospheric data remain limited. Second, comparative evaluations of different nonparametric estimators within the STSR framework are still scarce, particularly for environmental time-series data characterized by strong seasonality. Third, the potential of global reanalysis datasets such as MERRA-2 for semiparametric time-series modeling has not been extensively explored, despite their suitability for investigating nonlinear patterns and temporal dependencies. To address these gaps, this study develops relative humidity models using three different approaches: Fourier Series Nonparametric Regression (FSNR), Semiparametric Time Series Regression-Fourier Series (STSR-Fourier), and Semiparametric Time Series Regression-Local Polynomial (STSR-Local Polynomial). The proposed STSR models incorporate a first-order time-lag component as the parametric part to represent temporal dependence, while Fourier Series and Local Polynomial estimators are employed as nonparametric components to capture nonlinear and seasonal patterns. The analysis is conducted using monthly MERRA-2 relative humidity data from four observation stations in East Java, Indonesia, namely Juanda, Tanjung Perak, Karangploso, and Banyuwangi.

The main contributions of this study are threefold. First, it extends the application of the STSR framework to relative humidity data derived from the MERRA-2 reanalysis dataset. Second, it investigates the effectiveness of integrating time-lag effects with two different nonparametric estimators, namely Fourier Series and Local Polynomial estimators, for modeling hydrometeorological time-series data. Third, it provides a comprehensive comparative evaluation of FSNR, STSR-Fourier, and STSR-Local Polynomial models to identify the most effective approach for capturing RH dynamics. The findings are expected to contribute to the methodological development of semiparametric time-series regression and provide a flexible and accurate alternative for hydrometeorological data analysis. The remainder of this paper is

organized as follows. Section 2 presents the methodological framework of FSNR, STSR–Fourier, and STSR–Local Polynomial. Section 3 describes the study area, data, and research methodology. Section 4 presents the empirical results and discussion. Finally, Section 5 concludes the study.

2. PRELIMINARIES

2.1 Fourier Series Nonparametric Regression

Suppose that a set of paired observations (t, y_t) follows the Fourier Series Nonparametric Regression (FSNR) model given in Equation (1)

$$y_t = g(t) + \varepsilon_t, \quad (1)$$

where $t = 1, 2, \dots, n$.

The form of the regression function $g(t)$ is unknown and, in this study, it is estimated using the FSNR estimator. Assume that the regression function $g(t)$ belongs to the Hilbert space $L_2[a, b]$, i.e., $g(t) \in L_2[a, b]$. Therefore, $g(t)$ can be represented as a linear combination of the basis elements of $L_2[a, b]$ [20]. The estimation of the unknown Fourier series coefficients is carried out by determining the optimal value of λ . The optimal value of λ represents the number of Fourier coefficients (γ_w) that determine the smoothness of the estimated regression curve. Since the Fourier series employed in this study is a function of time and exhibits periodic behavior, the regression function $g(t)$ is estimated using a linear model involving sine and cosine functions [21], [22].

The complex exponential function can be expressed in terms of sine and cosine functions, which form a complete orthonormal system in $L_2[a, b]$, as given in Equation (2)

$$x_w(t) = e^{2\pi i w t} ; w = 0, \pm 1, \pm 2, \dots ; t = 1, 2, \dots, n ; i = \sqrt{-1}. \quad (2)$$

In the context of Fourier series, i denotes the imaginary unit.

The Fourier estimation function can be expressed in real form as a combination of cosine and sine functions. Therefore, the Fourier series estimator for $g(t)$ can be written as follows:

$$\hat{g}(t) = \frac{1}{2} \hat{\alpha}_0 + \hat{\gamma}t + \sum_{k=1}^{\lambda} \left(\hat{\alpha}_k \cos\left(\frac{2\pi k(t-1)}{n}\right) \right) + \left(\hat{\beta}_k \sin\left(\frac{2\pi k(t-1)}{n}\right) \right). \quad (3)$$

where $g(t)$ is the nonparametric regression function approximated using a Fourier series basis in a time series framework. The parameters $\hat{\alpha}_0$, $\hat{\gamma}$, $\hat{\alpha}_k$, and $\hat{\beta}_k$ are unknown regression coefficients estimated from the data. The parameter k indicates the oscillation order of the Fourier series, while t denotes the time predictor variable and y_t represents the response variable at the t -th observation [20], [23].

2.2 Semiparametric Time Series Regression

Semiparametric Time Series Regression (STSR) is a modeling technique that combines parametric and nonparametric components in time series data analysis [11]. This approach is employed when the relationship between predictor variables and the response variable is not completely known. Gao explained that several semiparametric time series regression models have been developed, one of which is the partially linear autoregressive model [24]. The form of a partially linear autoregressive model with lag 1, where the parametric component is represented by the lagged response variable (y_{t-1}), can be expressed as follows

$$y_t = \beta_0 + \beta_1 y_{t-1} + g(t) + \varepsilon_t, \quad (4)$$

where y_t denotes the response variable at time t , while y_{t-1} represents the value of the response variable at the previous time point (lag-1), which serves as the parametric component. The parameter β_0 is the intercept term, and β_1 is the autoregressive coefficient that measures the influence of past values on the current value. The function $g(t)$ is an unknown nonparametric component, with t acting as the predictor variable that is modeled flexibly to capture nonlinear patterns in the data. Furthermore, ε_t denotes the random error term associated with the time series observation at time t .

2.3 Semiparametric Time Series Regression - Fourier Series (STSR-Fourier)

Suppose we are given a time-series dataset consisting of a response variable y_t . The predictor variables include an autoregressive component y_{t-1} , and a time index t with $t = 2, 3, \dots, T$ following the STSR-Fourier model. Assume that the pattern of relationship between the predictor variables (y_{t-1}, t) and the response variable (y_t) follows a semiparametric regression model:

$$y_t = f(y_{t-1}) + g(t) + \varepsilon_t, \quad (5)$$

where ε_t is random error $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma^2$.

The parametric component is defined as $f(y_{t-1})$, which the Linear estimator will approximate, while the nonparametric component is defined as $g(t)$, which will be approximated by the complex exponential basis Fourier series estimator. The parametric component in Equation (5) is assumed to be known, namely linear, while the nonparametric component is unknown. The first thing in estimating the nonparametric component is to assume the parameters of the parametric component β are known with $y_t^* = y_t - \beta y_{t-1}$, so that:

$$y_t^* = g(t) + \varepsilon_t. \quad (6)$$

The nonparametric component $g(t)$ in Equation (6) is approximated by the complex exponential basis Fourier series estimator written in Equation (3). If t samples are taken at a time, then the complete STSR-Fourier model.

$$\begin{aligned} y_2 &= \beta_0 + \beta_1 y_1 + \frac{1}{2} \alpha_0 + \gamma t_2 + \sum_{k=1}^{\lambda} \left[\alpha_k \cos\left(\frac{2\pi k(t_2 - 1)}{n}\right) + \vartheta_k \sin\left(\frac{2\pi k(t_2 - 1)}{n}\right) \right] \\ y_3 &= \beta_0 + \beta_1 y_2 + \frac{1}{2} \alpha_0 + \gamma t_3 + \sum_{k=1}^{\lambda} \left[\alpha_k \cos\left(\frac{2\pi k(t_3 - 1)}{n}\right) + \vartheta_k \sin\left(\frac{2\pi k(t_3 - 1)}{n}\right) \right] \\ &\vdots \\ y_T &= \beta_0 + \beta_1 y_{T-1} + \frac{1}{2} \alpha_0 + \gamma t_T + \sum_{k=1}^{\lambda} \left[\alpha_k \cos\left(\frac{2\pi k(t_T - 1)}{n}\right) + \vartheta_k \sin\left(\frac{2\pi k(t_T - 1)}{n}\right) \right] \end{aligned}$$

So, it can be written in matrix notation in Equation (7)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (7)$$

where:

$$\mathbf{y} = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_T \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}; \quad \boldsymbol{\gamma} = \begin{bmatrix} \alpha_0 \\ \gamma \\ \alpha_1 \\ \vdots \\ \alpha_\lambda \\ \vartheta_1 \\ \vdots \\ \vartheta_\lambda \end{bmatrix}; \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_T \end{bmatrix};$$

$$\mathbf{X} = \begin{bmatrix} 1 & y_2 \\ 1 & y_3 \\ \vdots & \vdots \\ 1 & y_{T-1} \end{bmatrix}; \text{ and}$$

$$\mathbf{W} = \begin{bmatrix} 1/2 & t_2 & \cos\left(\frac{2\pi(t_2-1)}{n}\right) & \cdots & \cos\left(\frac{2\pi\lambda(t_2-1)}{n}\right) & \sin\left(\frac{2\pi(t_2-1)}{n}\right) & \cdots & \sin\left(\frac{2\pi\lambda(t_2-1)}{n}\right) \\ 1/2 & t_3 & \cos\left(\frac{2\pi(t_3-1)}{n}\right) & \cdots & \cos\left(\frac{2\pi\lambda(t_3-1)}{n}\right) & \sin\left(\frac{2\pi(t_3-1)}{n}\right) & \cdots & \sin\left(\frac{2\pi\lambda(t_3-1)}{n}\right) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1/2 & t_T & \cos\left(\frac{2\pi(t_T-1)}{n}\right) & \cdots & \cos\left(\frac{2\pi\lambda(t_T-1)}{n}\right) & \sin\left(\frac{2\pi(t_T-1)}{n}\right) & \cdots & \sin\left(\frac{2\pi\lambda(t_T-1)}{n}\right) \end{bmatrix}$$

The estimated value for the parameter vector $\hat{\boldsymbol{\gamma}}$ is obtained by partially differentiating Equation (7) with respect to $\boldsymbol{\gamma}$ and then sets it to zero. The LS estimator for the parameter $\hat{\boldsymbol{\gamma}}$ is written in Equation (8)

$$\hat{\boldsymbol{\gamma}} = (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \mathbf{y}^* . \quad (8)$$

The value of $\hat{\boldsymbol{\beta}}$ is obtained by partially differentiating Equation (7) with respect to $\boldsymbol{\beta}$ and then sets it to zero, so that it can be written in Equation (9)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F}) \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F}) \mathbf{y} . \quad (9)$$

Next, we substitute Equation (9) into Equation (7), such that we obtain

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{F}(\hat{\mathbf{y}} - \mathbf{X} \hat{\boldsymbol{\beta}}) \\ &= \mathbf{X}((\mathbf{X}^T (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F}) \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F}) \mathbf{y}) + \mathbf{F}(\hat{\mathbf{y}} - \mathbf{X}((\mathbf{X}^T (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F}) \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F}) \mathbf{y})) \end{aligned}$$

where:

$$\mathbf{B}_{par} = \mathbf{X}((\mathbf{X}^T (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F}) \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F})); \text{ and}$$

$$\mathbf{B}_{nonpar} = \mathbf{W}((\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T (\mathbf{I} - \mathbf{B}_{par})).$$

The general form of the prediction in the STSR-Fourier model, which is expressed in Equation (10).

$$\hat{\mathbf{y}} = (\mathbf{B}_{par} + \mathbf{B}_{nonpar}) \mathbf{y} \quad (10)$$

where $\mathbf{B}_{STSR-Fourier} = \mathbf{B}_{par} + \mathbf{B}_{nonpar}$.

The method used to determine the optimal number of oscillations in the STSR-Fourier model using

the Generalized Cross-Validation (GCV) method. The GCV formula for the STSR-Fourier model is written in Equation (11).

$$GCV(\lambda_{opt}) = \frac{T^{-1}(y^T(I - \mathbf{B}_{STSR-Fourier}))^T((I - \mathbf{B}_{STSR-Fourier})\mathbf{y})}{(T^{-1}tr(I - \mathbf{B}_{STSR-Fourier}))^2} \quad (11)$$

In Equation (11), n denotes the number of observations, and $tr(I - \mathbf{B}_{STSR-Fourier})$ represents the trace of the smoothing matrix $(I - \mathbf{B}_{STSR-Fourier})$, which reflects the model's level of complexity. The minimum value of GCV will provide the STSR-Fourier model with the optimal number of oscillations [19].

2.4 Semiparametric Time Series Regression - Local Polynomial (STSR-Local Polynomial)

Suppose that a time series dataset consists of a response variable y_t , an autoregressive variable y_{t-1} , and a time variable t , for $t = 2, 3, \dots, T$. The relationship between the response and predictor variables is assumed to follow the following semiparametric time series regression model:

$$y_t = \theta y_{t-1} + g(t) + \varepsilon_t \quad (12)$$

where y_t denotes the response variable at time t , y_{t-1} represents the first lag of the response variable, θ is the autoregressive parameter, $g(t)$ is an unknown nonparametric regression function, and ε_t is a random error term satisfying $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma^2$. The first step in the estimation procedure is to assume that the parameter θ is known. Subsequently, the partial response variable is defined as

$$y_t^* = y_t - \theta y_{t-1} . \quad (13)$$

Thus, Equation (13) can be rewritten as

$$y_t^* = g(t) + \varepsilon_t , \quad (14)$$

the function $g(t)$ is assumed to be smooth and possess derivatives up to a certain order. Therefore, it can be approximated using a Taylor series expansion around the point t_0 as follows:

$$g(t) \approx \sum_{d=0}^p \beta_d(t_0)(t_t - t_0)^d , \quad (15)$$

where p denotes the degree of the local polynomial and $\beta_d(t_0)$ represents the local parameters

to be estimated. Assuming that there are $n - 1$ observations, Equation (15) can be expressed in matrix form as

$$\mathbf{y}^* = \mathbf{V}(t_0)\boldsymbol{\beta}(t_0) + \boldsymbol{\varepsilon}, \quad (16)$$

where:

$$\mathbf{y}^* = \begin{bmatrix} y_2^* \\ y_3^* \\ \vdots \\ y_T^* \end{bmatrix};$$

$$\mathbf{V}(t_0) = \begin{bmatrix} 1 & (t_2 - t_0) & (t_2 - t_0)^2 & \cdots & (t_2 - t_0)^p \\ 1 & (t_3 - t_0) & (t_3 - t_0)^2 & \cdots & (t_3 - t_0)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (t_n - t_0) & (t_n - t_0)^2 & \cdots & (t_n - t_0)^p \end{bmatrix}.$$

The local parameter vector $\boldsymbol{\beta}(t_0)$ is estimated using the Weighted Least Squares (WLS) method by minimizing the objective function

$$Q(t_0) = (\mathbf{y}^* - \mathbf{V}(t_0)\boldsymbol{\beta}(t_0))^T \mathbf{K}_h (\mathbf{y}^* - \mathbf{V}(t_0)\boldsymbol{\beta}(t_0)), \quad (17)$$

where $\mathbf{K}_h$ is the kernel weighting matrix that depends on the bandwidth parameter h . By differentiating Equation (17) with respect to $\boldsymbol{\beta}(t_0)$ and setting the result equal to zero, the local polynomial estimator is obtained in Equation (18).

$$\hat{\boldsymbol{\beta}}(t_0) = (\mathbf{V}(t_0)^T \mathbf{K}_h \mathbf{V}(t_0))^{-1} \mathbf{V}(t_0)^T \mathbf{K}_h \mathbf{y}^*, \quad (18)$$

Substituting Equation (18) into Equation (15) yields the local polynomial estimator in the STSR model

$$\hat{\mathbf{g}}(t) = \mathbf{A}(t_0)\mathbf{y}^*,$$

where $\mathbf{A}(t_0) = \mathbf{V}(t_0)(\mathbf{V}(t_0)^T \mathbf{K}_h \mathbf{V}(t_0))^{-1} \mathbf{V}(t_0)^T \mathbf{K}_h \mathbf{V}(t_0)$.

Which is referred to as the local polynomial smoothing matrix.

Next, the autoregressive parameter θ associated with the parametric component is estimated using the Least Squares method by minimizing the sum of squared errors. The estimator of the autoregressive parameter is given in Equation (19)

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{y}_{t-1}^T (\mathbf{I} - \mathbf{A}(t_0))^T (\mathbf{I} - \mathbf{A}(t_0)) \mathbf{y}_{t-1} \right)^{-1} \mathbf{y}_{t-1}^T (\mathbf{I} - \mathbf{A}(t_0))^T (\mathbf{I} - \mathbf{A}(t_0)) \mathbf{y}_t. \quad (19)$$

Subsequently, the estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{g}}(t)$ are substituted into the model, yielding the estimated

STSR–Local Polynomial model:

$$\begin{aligned}\hat{y}_t &= \hat{\theta}y_{t-1} + \hat{g}(t) \\ &= \left(\left(y_{t-1}^T (\mathbf{I} - \mathbf{A}(t_0))^T (\mathbf{I} - \mathbf{A}(t_0)) y_{t-1} \right)^{-1} y_{t-1}^T (\mathbf{I} - \mathbf{A}(t_0))^T (\mathbf{I} - \mathbf{A}(t_0)) y_t \right) y_{t-1} + \\ &\quad \mathbf{A}(t_0) \left(y_t - \left(\left(y_{t-1}^T (\mathbf{I} - \mathbf{A}(t_0))^T (\mathbf{I} - \mathbf{A}(t_0)) y_{t-1} \right)^{-1} y_{t-1}^T (\mathbf{I} - \mathbf{A}(t_0))^T (\mathbf{I} - \right. \right. \\ &\quad \left. \left. \mathbf{A}(t_0)) y_t \right) y_{t-1} \right).\end{aligned}\tag{20}$$

$$\text{Let } \mathbf{C} = y_{t-1} \left(\left(y_{t-1}^T (\mathbf{I} - \mathbf{A}(t_0))^T (\mathbf{I} - \mathbf{A}(t_0)) y_{t-1} \right)^{-1} y_{t-1}^T (\mathbf{I} - \mathbf{A}(t_0))^T (\mathbf{I} - \mathbf{A}(t_0)) \right),$$

then Equation (20) can be rewritten as $\hat{y}_t = \mathbf{C}y_t + \mathbf{A}(t_0)(\mathbf{I} - \mathbf{C})y_t$,

where $\mathbf{A}(t_0) = [\mathbf{a}_{(2)t_0}, \mathbf{a}_{(3)t_0}, \dots, \mathbf{a}_{(T)t_0}]^T$ and $\mathbf{a}_{(T)t_0} = \mathbf{v}_{(t)t_0} (\mathbf{V}_{t_0}^T \mathbf{K}_h(t_0) \mathbf{V}_{t_0})^{-1} \mathbf{V}_{t_0}^T \mathbf{K}_h(t_0)$

The performance of the local polynomial estimator is highly influenced by the choice of bandwidth.

The optimal bandwidth is determined using the GCV criterion, defined in Equation (21).

$$GCV(h_{opt}) = \frac{T^{-1} (y^T (\mathbf{I} - \mathbf{A}_h))^T ((\mathbf{I} - \mathbf{A}_h) y)}{(T^{-1} \text{tr}(\mathbf{I} - \mathbf{A}_h))^2}, \tag{21}$$

where $\mathbf{A}_h = \mathbf{V}(t_0) (\mathbf{V}(t_0)^T \mathbf{K}_h \mathbf{V}(t_0))^{-1} \mathbf{V}(t_0)^T \mathbf{K}_h \mathbf{V}(t_0)$.

The bandwidth corresponding to the minimum GCV value is selected as the optimal bandwidth because it provides the best compromise between estimation bias and variance. Consequently, the resulting estimated curve is neither excessively rough nor overly smooth, allowing it to represent the underlying data pattern more accurately [16], [25], [26].

2.5 Accuracy Metrics

Model evaluation is conducted to assess the ability of the model to represent the underlying data pattern and generate accurate predictions [27], [28]. In this study, model performance is evaluated using four prediction error measures, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Normalized Mean Squared Error (NMSE) and Mean Absolute Percentage Error (MAPE) [29]. These measures are employed simultaneously because each provides a different perspective on prediction accuracy and model performance.

The MSE is defined in Equation (22)

$$MSE = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2 . \quad (22)$$

The RMSE is defined in Equation (23)

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2} . \quad (23)$$

The NMSE is defined in Equation (24)

$$NMSE = \frac{\sum_{t=1}^T (y_t - \hat{y}_t)^2}{\sum_{t=1}^T (y_t - \bar{y})^2} . \quad (24)$$

The MAPE is defined in Equation (25)

$$MAPE = \frac{100}{n} \sum_{t=1}^T \left| \frac{y_t - \hat{y}_t}{y_t} \right| . \quad (25)$$

Lower values of MSE, RMSE, NMSE, and MAPE indicate better predictive performance. MAPE provides an easily interpretable measure of prediction accuracy because it expresses forecasting errors as percentages. A MAPE value below 10% indicates highly accurate forecasting, a value between 10% and 20% indicates good forecasting performance, a value between 20% and 50% indicates reasonable forecasting accuracy, while a value exceeding 50% suggests poor forecasting performance [30].

2.6 Research Methodology

2.6.1 Data and Data Sources

The data used in this study were obtained from the NASA Power Study MERRA-2 (Modern-Era Retrospective Analysis for Research and Applications, Version 2) reanalysis dataset <https://power.larc.nasa.gov/data-access-viewer/>. MERRA-2 is a global reanalysis product that assimilates multiple observations, including satellite, ground-based, and sounding measurements, into physical models, providing comprehensive atmospheric and environmental data.

For this study, we specifically used the Relative Humidity (RH2M), Precipitation Corrected Sum (PRECTOTCORR_SUM), Earth Skin Temperature (TS) datasets. The observation period spans January 2015–December 2025, with data extracted from the grid points closest to the study

sites.

Table 1. Location and Elevation of Study Sites

Station	Coordinate	Elevation from MERRA-2 (m)
Juanda	Location: Latitude -7.38 Longitude 112.7864	Average for 0.5 x 0.625 degree lat/lon region = 239.03 meters
Perak	Location: Latitude -7.1969 Longitude 112.7324	Average for 0.5 x 0.625 degree lat/lon region = 18.05 meters
Blimbingsari	Location: Latitude -8.3105 Longitude 114.3403	Average for 0.5 x 0.625 degree lat/lon region = 106.56 meters
Karangploso	Location: Latitude -7.8971 Longitude 112.5997	Average for 0.5 x 0.625 degree lat/lon region = 467.71 meters

The four BMKG stations were selected to represent relative humidity variability in East Java Province, particularly in industrial and rapidly urbanizing areas. These locations capture differences in elevation and coastal–inland characteristics, allowing the analysis to reflect spatial heterogeneity in humidity behavior. Moreover, East Java is highly relevant for climate change studies, as industrial development and changing climatic conditions are expected to influence atmospheric moisture dynamics.

2.6.2 Framework Conceptual

The research flowchart is used to provide a systematic overview of the research stages, beginning with data collection and ending with model evaluation and selection. The proposed framework was designed to compare the performance of nonparametric regression and semiparametric time series regression models in modeling Relative Humidity data observed at several hydrological monitoring stations. The research workflow is illustrated in Figure (1).

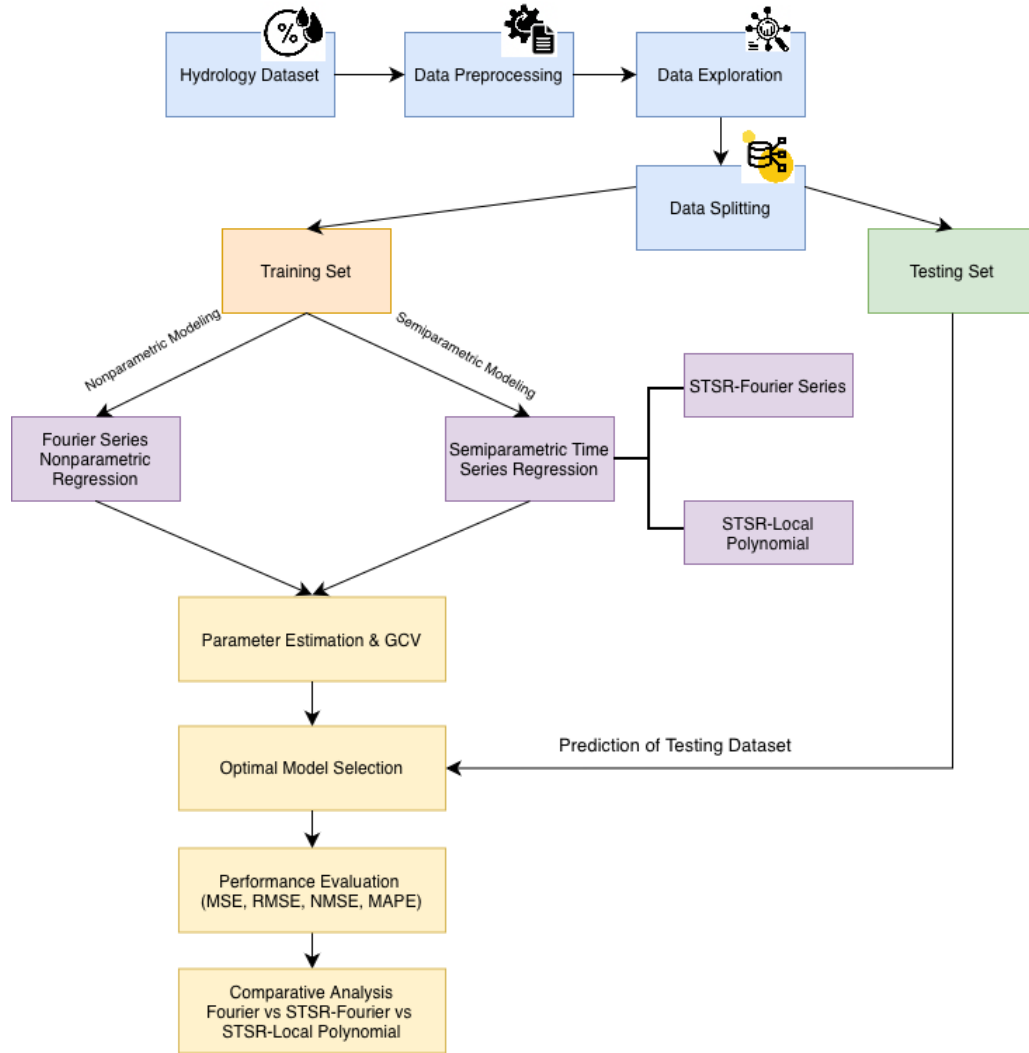


Figure 1. Research Framework

All computational procedures in this study were implemented using RStudio with the R programming language. The implementation consisted of data preprocessing, exploratory data analysis, model construction, parameter estimation, optimal parameter selection based on the Generalized Cross-Validation (GCV) criterion, prediction, and model performance evaluation. The FSNR, STSR–Fourier, and STSR–Local Polynomial models were developed using customized scripts based on the proposed methodological framework. The computational procedures for each model are summarized in Algorithms 1–3.

Algorithm 1. Fourier Series Nonparametric Regression (FSNR)

Input: Response variable, time index, maximum number of oscillations, and training proportion.

1. Split the dataset into training and testing sets.
2. For each number of oscillations:
 - a. Estimate the Fourier coefficients.
 - b. Construct the Fourier Series model.
 - c. Calculate the GCV value.
3. Select the number of oscillations with the minimum GCV value.
4. Re-estimate the Fourier Series model using the selected oscillation.
5. Generate predictions for the training and testing datasets.
6. Calculate MSE, RMSE, NMSE, and MAPE.

Output: Optimal Fourier Series model, predicted values, and performance measures.

Algorithm 2. Semiparametric Time Series Regression–Fourier Series (STSR–Fourier)

Input: Response variable, time-lag variable, time index, maximum number of oscillations, and training proportion.

1. Split the dataset into training and testing sets.
2. For each number of oscillations ($k = 1, 2, \dots, 10$):
 - a. Construct the Fourier Series basis functions using the time index.
 - b. Construct the STSR–Fourier model using the time-lag variable as the parametric component and the Fourier Series basis functions as the nonparametric component.
 - c. Estimate the model parameters.
 - d. Calculate the GCV value.
3. Select the number of oscillations with the minimum GCV value.
4. Re-estimate the STSR–Fourier model using the selected oscillation.
5. Generate predictions for the training and testing datasets.
6. Calculate MSE, RMSE, NMSE, and MAPE.

Output: Optimal STSR–Fourier model, predicted values, and performance measures.

Algorithm 3. Semiparametric Time Series Regression–Local Polynomial (STSR–Local Polynomial)

Input: Response variable, time-lag variable, time index, bandwidth values ranging from 0.1 to 10 with an increment of 0.89, and training proportion.

1. Split the dataset into training and testing sets.
2. For each bandwidth value ($h = 0.1, 0.99, 1.88, \dots, 10$):
 - a. Construct the first-order Local Polynomial estimator using the time index.
 - b. Construct the STSR–Local Polynomial model using the time-lag variable as the parametric component and the first-order Local Polynomial estimator as the nonparametric component.
 - c. Estimate the model parameters.
 - d. Calculate the GCV value.
3. Select the bandwidth with the minimum GCV value.
4. Re-estimate the STSR–Local Polynomial model using the selected bandwidth.
5. Generate predictions for the training and testing datasets.
6. Calculate MSE, RMSE, NMSE, and MAPE.

Output: Optimal STSR–Local Polynomial model, predicted values, and performance measures.

3. MAIN RESULTS

3.1 Time Series Data Exploration

Time series exploration was conducted as a preliminary step to understand the characteristics of the Relative Humidity (RH) data used in this study. This stage aims to identify underlying patterns in the data, including trends, seasonal fluctuations, and potential temporal dependencies. The monthly RH time series for each observation station is presented in Figure 2.

TIME-LAG SEMIPARAMETRIC MODEL OF RELATIVE HUMIDITY

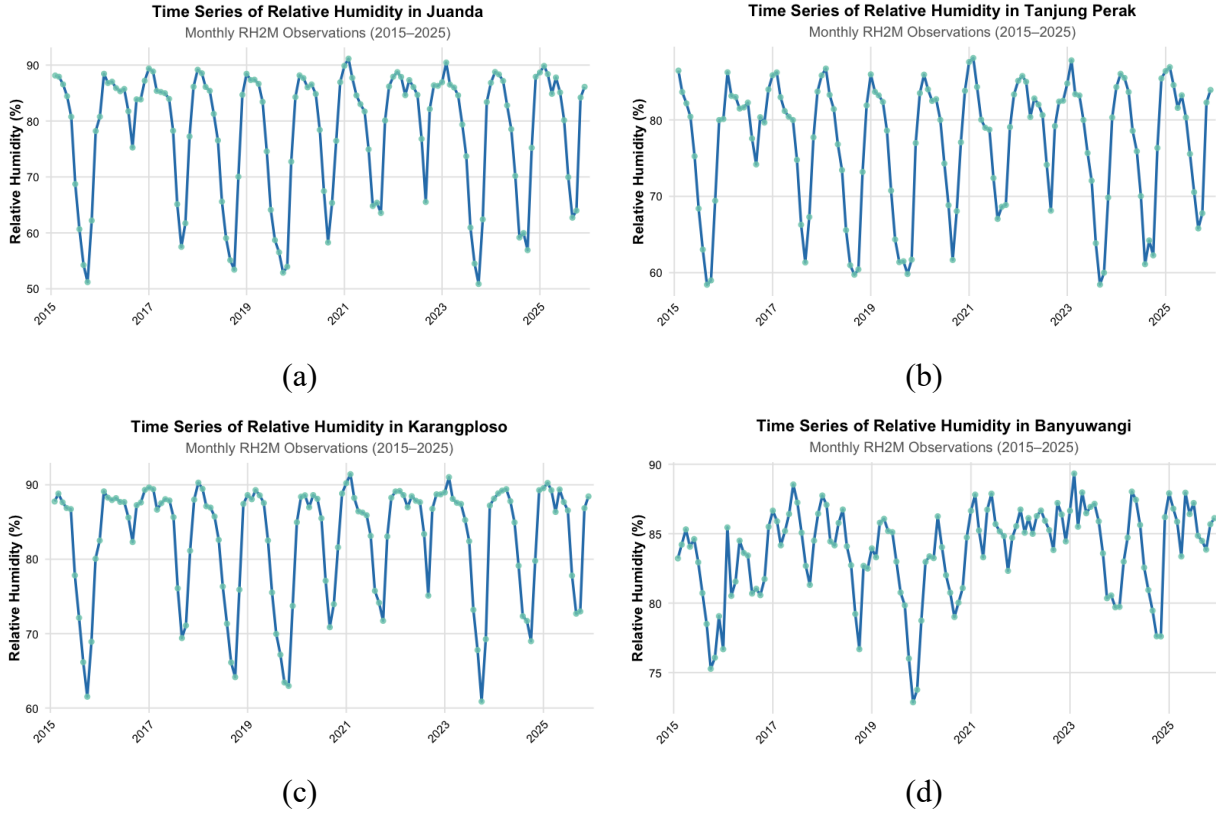


Figure 2. Time series plots of monthly relative humidity at (a) Juanda, (b) Tanjung Perak, (c) Karangploso, and (d) Banyuwangi during 2015–2025

Figures 2 illustrate the monthly Relative Humidity (RH) observations recorded at Juanda, Tanjung Perak, Karangploso, and Banyuwangi during the period 2015-2025. In general, all locations exhibit a pronounced annual seasonal pattern characterized by recurring fluctuations from year to year. This recurring behavior indicates that atmospheric humidity in East Java is strongly influenced by the regional monsoon climate system, which governs the alternation between wet and dry seasons. At Juanda, RH values fluctuate within a relatively wide range, with consistent annual peaks and troughs. Humidity levels tend to increase during the wet season and reach their highest values between the end and beginning of each year. Conversely, RH decreases substantially during the middle of the year, coinciding with the dry season. A similar pattern is observed at Tanjung Perak. Given their geographical proximity and coastal characteristics, both stations exhibit comparable seasonal dynamics, including relatively large fluctuation amplitudes

and sharp decreases in humidity during dry-season months.

Karangploso exhibits somewhat different characteristics. Located in a higher-altitude area, this station generally maintains higher RH levels throughout the observation period. Although seasonal fluctuations remain evident, the variation in RH is comparatively smaller than that observed at Juanda and Tanjung Perak. This finding suggests that topographic conditions contribute to maintaining relatively stable atmospheric moisture throughout the year. In contrast, Banyuwangi displays a more moderate pattern of RH variability. Monthly fluctuations are generally smaller, and the seasonal pattern appears less pronounced than at the other stations. Additionally, there is a slight indication of increasing average RH levels during the later years of the observation period. However, further statistical analysis is required to determine whether this pattern represents a significant long-term trend.

Overall, RH values across all stations tend to reach their maximum levels during the wet season, particularly between December and March, while minimum values are generally observed during the dry season, especially in August and September. The clear seasonal structure observed in all locations suggests that periodic modeling approaches, such as Fourier-based methods, may be appropriate for capturing the temporal dynamics of RH data. Furthermore, the persistence of humidity levels across consecutive observations indicates the potential presence of temporal dependence, supporting the use of semiparametric time series regression models incorporating autoregressive components. A more detailed examination of monthly humidity variability is presented through the Monthly Distribution of Relative Humidity shown in Figure 6.

RH values across all observation stations generally reach their highest levels during the wet season, particularly between December and March, whereas the lowest values are typically observed during the dry season, especially in August and September. A more detailed illustration of the seasonal variation in Relative Humidity (RH) is presented through the monthly distributions shown in Figure 3.

TIME-LAG SEMIPARAMETRIC MODEL OF RELATIVE HUMIDITY

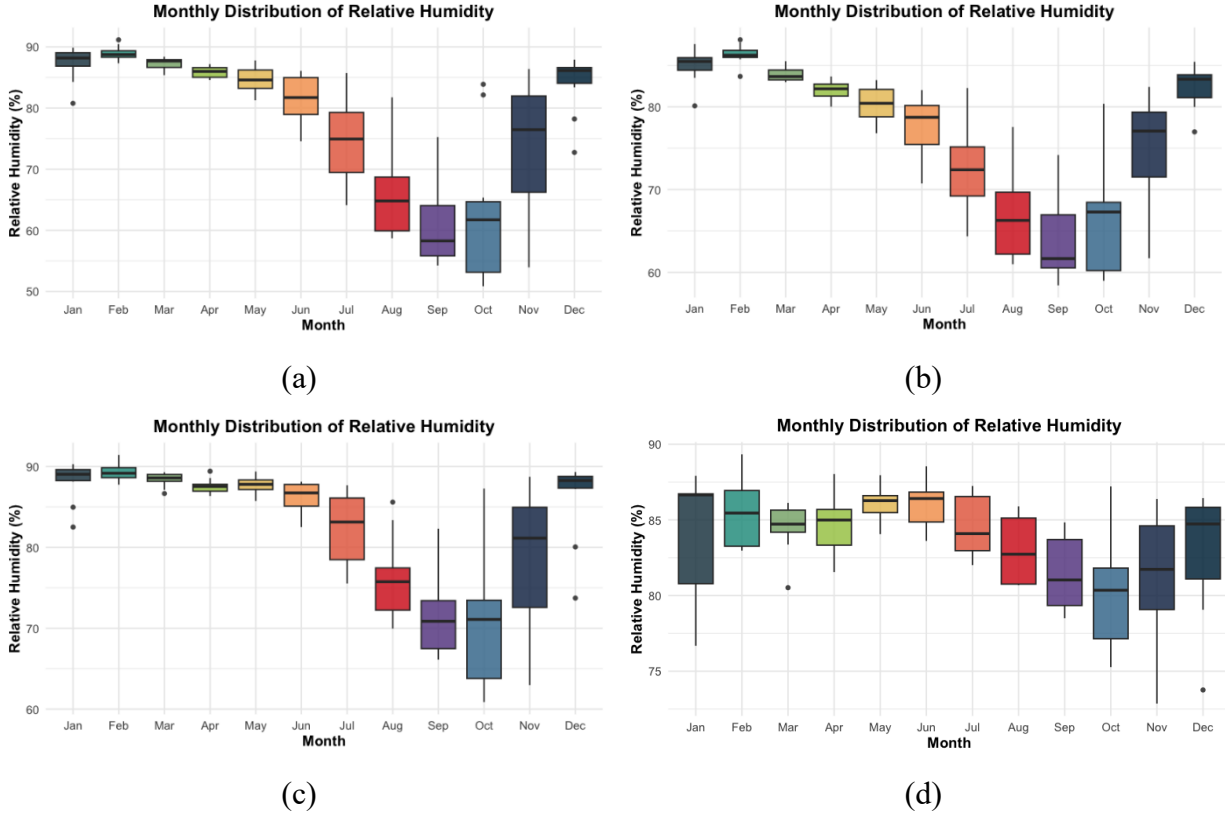


Figure 3. Monthly boxplots of relative humidity at (a) Juanda, (b) Tanjung Perak, (c) Karangploso, and (d) Banyuwangi during 2015–2025

Overall, the exploratory analysis of the time series plots, and monthly boxplots reveals a pronounced annual seasonal pattern across all observation stations. Relative humidity increases during the wet season and decreases during the dry season, exhibiting a recurring cycle that remains consistent throughout the observation period.

3.2 Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

Before constructing the semiparametric time series regression models, an analysis of the temporal dependence structure was conducted using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). This analysis aims to identify the presence of relationships between current observations and observations from previous time periods (time lags), thereby determining the necessity of incorporating autoregressive components into the model.

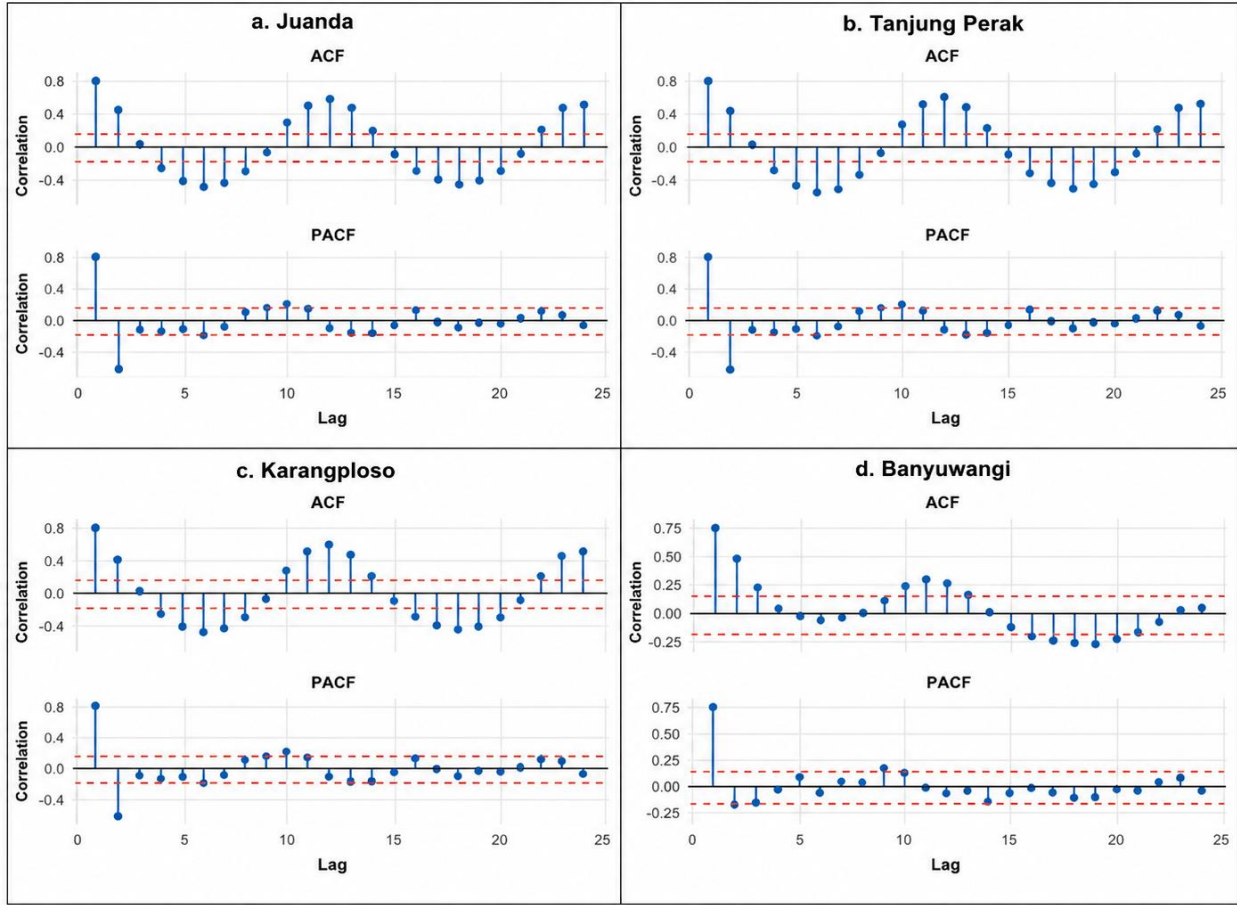


Figure 4. ACF and PACF plots of monthly relative humidity at (a) Juanda, (b) Tanjung Perak, (c) Karangploso, and (d) Banyuwangi

Based on the PACF in Figure 4, the first-order autoregressive component, y_{t-1} , was selected as the predictor in the parametric part of the Semiparametric Time Series Regression models. The first lag was chosen because it consistently exhibited a significant direct effect on the current Relative Humidity (RH) values across all observation stations. In contrast, the influence of higher-order lags generally diminished after controlling for the effect of the first lag. Therefore, the temporal dependence structure of the RH data can be adequately represented by an AR (1) component.

3.3 Scatter Plot

Scatter plots between y_{t-1} and y_t were constructed to provide preliminary justification for selecting the lag-1 variable as the parametric component in the STSR models. The observed

relationship exhibits a generally linear pattern, indicating that a portion of the current RH variability can be explained by the RH value from the previous period. Therefore, the inclusion of a first-order autoregressive component (AR(1)) is appropriate for capturing the temporal dependence structure present in the data.

In addition, scatter plots between time t and relative humidity y_t were used to examine the temporal behavior of RH. The plots exhibit a nonlinear and recurring seasonal pattern, indicating that the relationship cannot be adequately described by a simple parametric function. Therefore, a nonparametric approach is adopted to flexibly capture the temporal variation of RH. The scatter plots for all observation stations are presented in Figures 5–8.

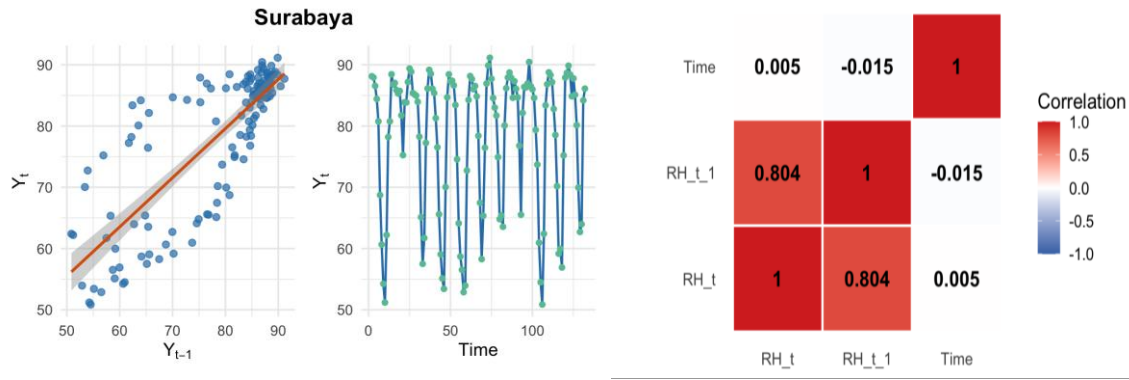


Figure 5. Lagged relationship and temporal structure of monthly relative humidity in Surabaya

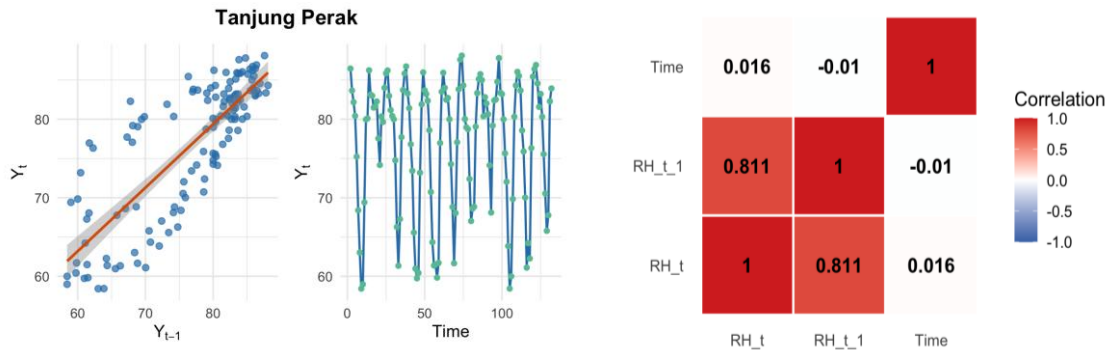


Figure 6. Lagged relationship and temporal structure of monthly relative humidity in Tanjung Perak

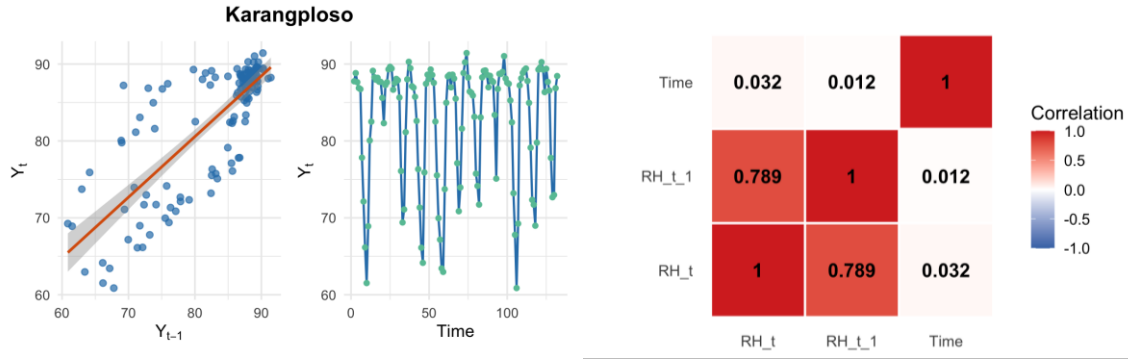


Figure 7. Lagged relationship and temporal structure of monthly relative humidity in Karangploso

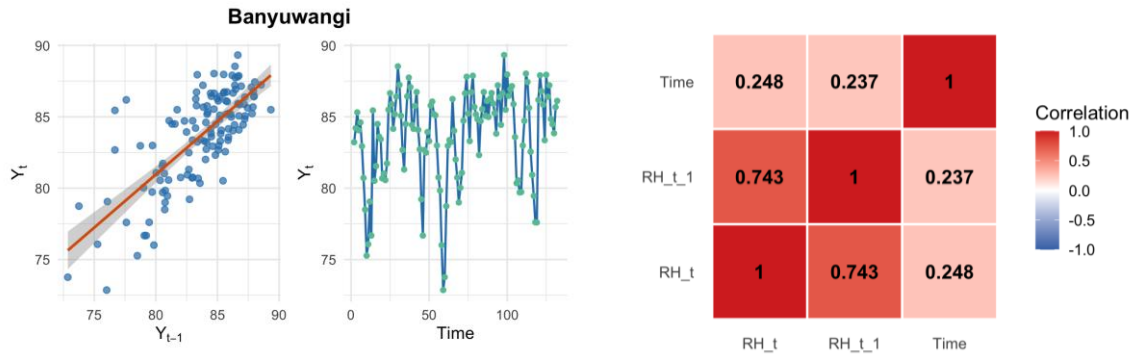


Figure 8. Lagged relationship and temporal structure of monthly relative humidity in Banyuwangi

Figures 5-8 show that the correlation between y_t and y_{t-1} is relatively strong across all observation stations, with correlation coefficients of 0.804 for Juanda, 0.811 for Tanjung Perak, 0.743 for Banyuwangi, and 0.789 for Karangploso. These results indicate the presence of substantial temporal dependence in the RH data, suggesting that y_{t-1} is an appropriate predictor for the parametric component of the STSR models. In contrast, the correlation between time t and y_t is relatively weak, implying that the relationship between time and RH cannot be adequately represented by a simple linear function. Therefore, a nonparametric component is required to capture the seasonal patterns and complex temporal dynamics of Relative Humidity.

3.4 Model Performance Evaluation

Prior to model development, the dataset was divided into training and testing subsets using three partitioning schemes, namely 95:5, 90:10, and 85:15. These partitioning strategies were employed to evaluate the stability and predictive performance of the proposed models under different proportions of training and testing data. The results of the FSNR, STSR-Fourier, and STSR-Local Polynomial models are presented in Tables 2-4, respectively.

Table 2. Performance of the FSNR Model

Observation Station	Scheme	Oscillation	Training					Testing			
			GCV	MSE	RMSE	NMSE	MAPE (%)	MSE	RMSE	NMSE	MAPE (%)
Juanda	95:5	11	38.257	25.381	5.038	0.189	5.597	118.161	10.870	1.319	11.428
	90:10	10	42.371	28.526	5.341	0.208	5.983	142.678	11.945	1.761	11.916
	85:15	10	31.977	21.022	4.585	0.158	5.143	334.319	18.284	2.719	22.005
Tanjung Perak	95:5	11	18.713	12.415	3.523	0.170	3.826	42.005	6.481	0.925	7.549
	90:10	10	21.045	14.168	3.764	0.191	4.130	66.465	8.153	1.454	8.457
	85:15	10	14.400	9.467	3.077	0.131	3.407	147.553	12.147	2.130	14.351
Karangploso	95:5	11	21.882	14.517	3.810	0.229	3.868	64.076	8.005	1.476	7.754
	90:10	10	23.016	15.495	3.936	0.239	4.091	78.601	8.866	2.177	8.443
	85:15	10	19.624	12.901	3.592	0.201	3.659	165.107	12.849	3.220	13.873
Banyuwangi	95:5	10	5.992	4.134	2.033	0.377	1.970	17.478	4.181	14.647	3.476
	90:10	10	4.956	3.336	1.827	0.309	1.805	27.911	5.283	4.395	5.854
	85:15	9	4.420	3.036	1.742	0.285	1.724	44.287	6.655	4.496	6.942

Table 2 shows that the FSNR model achieved satisfactory predictive performance across all observation stations, with the optimal number of oscillations ranging from 9 to 11. For Juanda, Tanjung Perak, Karangploso, and Banyuwangi, the 95:5 training–testing scheme consistently produced lower testing MSE, RMSE, NMSE, and MAPE values than the 90:10 and 85:15 schemes, indicating good predictive accuracy under this data partitioning strategy.

Table 3. Performance of the STSR-Fourier Series Model

Observation Station	Scheme	Oscillation	Training					Testing			
			GCV	MSE	RMSE	NMSE	MAPE (%)	MSE	RMSE	NMSE	MAPE (%)
Juanda	95:5	1	48.991	45.119	6.717	0.336	7.215	74.461	8.629	0.831	9.297
	90:10	1	48.060	44.040	6.636	0.322	7.138	86.801	9.317	1.072	8.750
	85:15	1	48.489	44.219	6.650	0.331	7.103	62.232	7.889	0.506	8.655
Tanjung Perak	95:5	1	25.973	23.921	4.891	0.327	5.284	37.049	6.087	0.816	6.619
	90:10	1	25.303	23.186	4.815	0.313	5.254	45.652	6.757	0.999	6.262
	85:15	1	25.219	22.998	4.796	0.318	5.178	35.043	5.920	0.506	6.181
Karangploso	95:5	1	24.749	22.793	4.774	0.360	4.629	36.858	6.071	0.849	5.853
	90:10	1	24.641	22.580	4.752	0.348	4.580	39.022	6.247	1.081	5.698
	85:15	1	25.054	22.848	4.780	0.356	4.583	28.391	5.328	0.554	5.386
Banyuwangi	95:5	2	5.311	4.891	2.212	0.446	2.140	1.244	1.115	1.043	1.189
	90:10	2	4.756	4.204	2.050	0.390	1.996	16.679	4.084	2.626	3.953
	85:15	1	4.919	4.486	2.118	0.421	2.059	6.469	2.543	0.657	2.530

According to Table 3, the STSR–Fourier model demonstrated consistent predictive performance across all observation stations. For Juanda, the best performance was achieved under the 85:15 training–testing scheme, with a testing RMSE of 7.889, NMSE of 0.506, and MAPE of 8.66%, which were lower than those obtained under the 95:5 and 90:10 schemes. Similarly, the best performance for Tanjung Perak was obtained under the 85:15 partition, yielding a testing RMSE of 5.920, NMSE of 0.506, and MAPE of 6.18%. The same pattern was observed for Karangploso, where the 85:15 scheme resulted in a testing RMSE of 5.328, NMSE of 0.554, and MAPE of 5.39%. In contrast, Banyuwangi achieved its best predictive performance under the 95:5 training–testing scheme, with a testing RMSE of 1.115 and MAPE of 1.19%, which were considerably lower than those obtained under the alternative partitioning schemes.

Table 4. Performance of the STSR–Local Polynomial Model

Observation Station	Scheme	Bandwidth	Training					Testing			
			GCV	MSE	RMSE	NMSE	MAPE (%)	MSE	RMSE	NMSE	MAPE (%)
Juanda	95:5	9.890	57.621	50.042	7.074	0.370	7.281	291.880	17.085	2.792	18.236
	90:10	9.890	57.921	50.113	7.079	0.363	7.278	902.715	30.045	10.348	35.940
	85:15	9.890	53.824	47.237	6.873	0.351	7.204	448.919	21.188	3.469	26.909
Tanjung Perak	95:5	9.890	31.257	27.046	5.201	0.366	5.456	208.429	14.437	3.935	16.660
	90:10	9.890	31.116	26.820	5.179	0.358	5.500	682.742	26.129	13.867	32.216
	85:15	9.890	28.877	25.155	5.015	0.345	5.356	404.426	20.110	5.545	26.007
Karangploso	95:5	9.890	28.838	25.111	5.011	0.393	4.672	267.322	16.350	5.278	17.874
	90:10	9.890	29.720	25.782	5.078	0.394	4.662	1012.736	31.824	26.051	37.057
	85:15	9.890	27.606	24.256	4.925	0.374	4.649	366.401	19.142	6.788	22.625
Banyuwangi	95:5	9.890	6.382	5.556	2.357	0.503	2.257	97.972	9.898	70.374	11.388
	90:10	9.890	5.976	5.235	2.288	0.481	2.157	601.135	24.518	87.887	28.548
	85:15	9.890	5.800	5.078	2.254	0.472	2.144	45.723	6.762	4.410	7.691

According to Table 4, the STSR–Local Polynomial model exhibited varying predictive performance across different data partitioning schemes. For each observation station and data partitioning scheme, the optimal bandwidth selected using the GCV criterion was 9.890. This result indicates that the same smoothing level was consistently identified as optimal across all model configurations. For Juanda, the best performance was achieved under the 95:5 training–testing scheme, with a testing RMSE of 17.08, NMSE of 2.79, and MAPE of 18.24%, which were

lower than those obtained under the 90:10 and 85:15 schemes. A similar pattern was observed for Tanjung Perak, where the 95:5 partition yielded a testing RMSE of 14.44, NMSE of 3.93, and MAPE of 16.66%. Likewise, Karangploso achieved its best performance under the 95:5 scheme, with a testing RMSE of 16.35, NMSE of 5.28, and MAPE of 17.87%. In contrast, Banyuwangi obtained its best predictive performance under the 85:15 partition, yielding a testing RMSE of 6.76, NMSE of 4.41, and MAPE of 7.69%.

3.5 Selection of Optimal Data Splitting Scheme

To establish an objective basis for model comparison, the data splitting scheme was selected using a majority voting approach based on the lowest testing MAPE obtained at each observation station. This approach was chosen because it accounts for the consistency of model performance across different locations. The results of the majority voting procedure are presented in Figure 9.

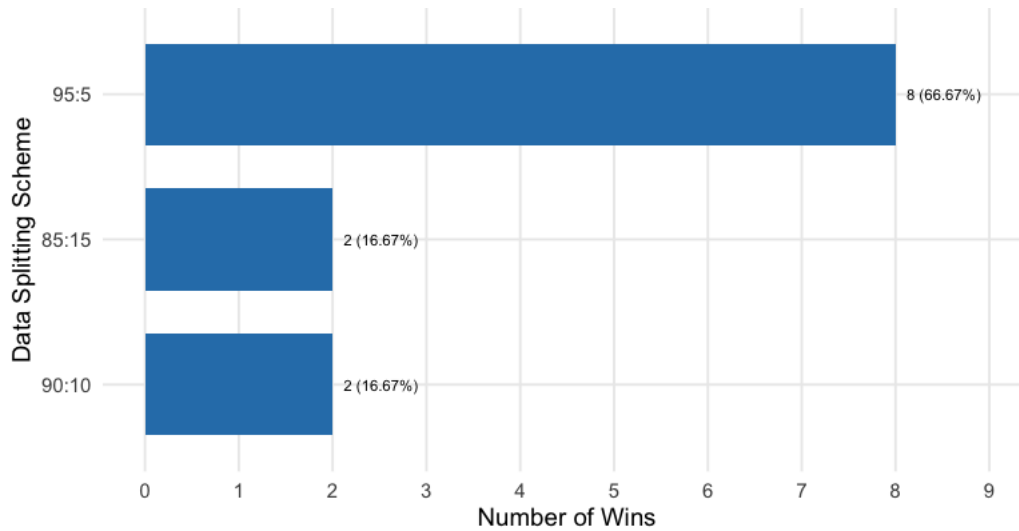


Figure 9. Selection of the Optimal Data Splitting Scheme Using Majority Voting

The results indicate that the 95:5 training–testing scheme achieved the highest number of wins, accounting for 8 out of 12 evaluation cases (66.67%). In contrast, the 90:10 and 85:15 schemes each obtained only 2 wins (16.67%). These findings suggest that the 95:5 partition provides the most consistent predictive performance across the majority of models and observation stations. Therefore, the 95:5 training–testing scheme was selected as the benchmark data partition for the subsequent comparison of model performance.

3.6 Comparative Analysis of Fourier, STSR-Fourier, and STSR-Local Polynomial

Based on the majority voting results, the 95:5 training–testing scheme was selected as the optimal data partition because it demonstrated the most consistent performance across the majority of models and observation stations. Therefore, all subsequent model comparisons were conducted using the testing results obtained under the 95:5 partition to ensure a fair and objective evaluation.

Figure 10 presents a comparison of the predictive performance of the FSNR, STSR–Fourier, and STSR-Local Polynomial models based on the testing data from the 95:5 partition. In general, the STSR-Fourier model produced lower prediction errors than the other two models across all observation stations. At Juanda, the MAPE decreased from 11.43% for the FSNR model to 9.30% for the STSR-Fourier model. Similar improvements were observed at Tanjung Perak and Karangploso, where the STSR-Fourier model achieved MAPE values of 6.62% and 5.85%, respectively, compared with 7.55% and 7.75% obtained by the FSNR model.

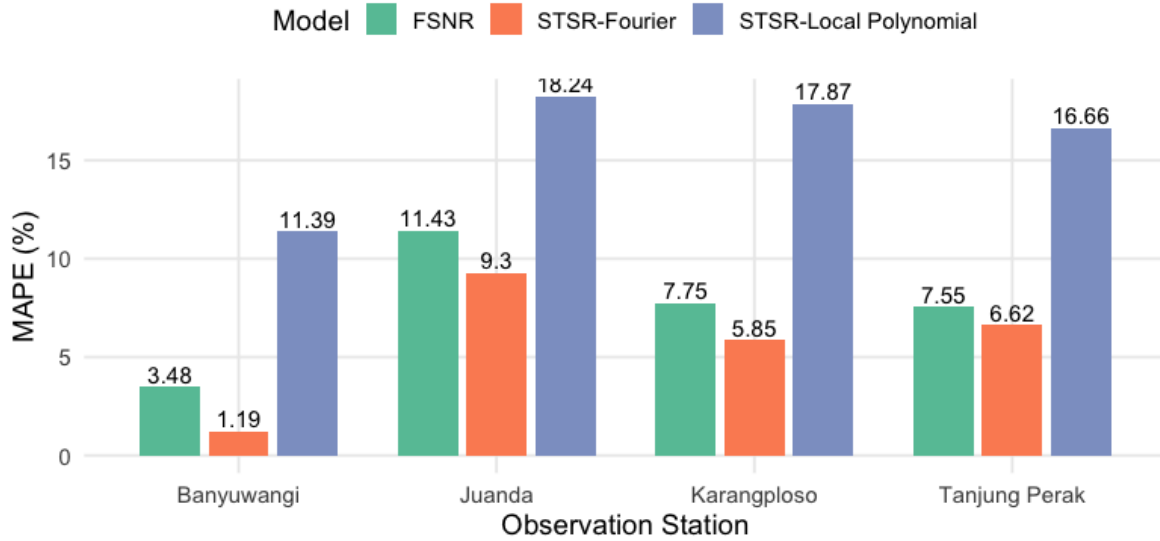


Figure 10. Comparison of Model Performance Based on Testing MAPE Under the 95:5 Training–Testing Scheme

The optimal STSR-Fourier model for Juanda is given in Equation (26)

$$\hat{y}_t = -5.264 \times 10^{-19} + 0.814y_{t-1} + 27.342 + 0.013t + 0.228 \cos\left(\frac{2\pi(t-1)}{n}\right) + 0.302 \sin\left(\frac{2\pi(t-1)}{n}\right). \quad (26)$$

Based on Equation (26), which represents the best-performing STSR-Fourier model for Juanda, the autoregressive coefficient of 0.814 indicates that Relative Humidity (RH) in the previous period has a strong positive influence on the current RH value. Furthermore, the Fourier component, consisting of sine and cosine functions, indicates the presence of periodic patterns successfully captured by the model. The positive time coefficient of 0.013 suggests a slight increasing tendency in RH over time, although its effect is relatively small compared with that of the autoregressive component.

The optimal STSR-Fourier model for Tanjung Perak is given in Equation (27)

$$\hat{y}_t = -3.300 \times 10^{-18} + 0.818y_{t-1} + 26.523 + 0.011t + 0.137 \cos\left(\frac{2\pi(t-1)}{n}\right) + 0.267 \sin\left(\frac{2\pi(t-1)}{n}\right). \quad (27)$$

Based on Equation (27), which represents the best-performing STSR-Fourier model for Tanjung Perak, the autoregressive coefficient of 0.818 indicates that RH in the previous period exerts a strong positive influence on the current RH value. In addition, the Fourier component, represented by sine and cosine functions, confirms the existence of periodic behavior captured by the model. Meanwhile, the time coefficient of 0.011 suggests only a minor temporal trend throughout the observation period.

The optimal STSR-Fourier model for Karangploso is given in Equation (28)

$$\hat{y}_t = 3.879 \times 10^{-18} + 0.799y_{t-1} + 32.080 + 0.010t + 0.170 \cos\left(\frac{2\pi(t-1)}{n}\right) + 0.183 \sin\left(\frac{2\pi(t-1)}{n}\right). \quad (28)$$

Based on Equation (28), which represents the best-performing STSR-Fourier model for Karangploso, the autoregressive coefficient of 0.799 indicates that RH in the previous period has a strong positive effect on the current RH value. The Fourier component, consisting of sine and cosine functions, captures the recurring seasonal fluctuations present in the data. Furthermore, the time coefficient of 0.010 indicates only a slight temporal variation throughout the study period.

The optimal STSR-Fourier model for Banyuwangi is given in Equation (29)

$$\hat{y}_t = -2.124 \times 10^{-17} + 0.686y_{t-1} + 49.517 + 0.024t + 0.059 \cos\left(\frac{2\pi(t-1)}{n}\right) + \quad (29)$$

$$0.677 \sin\left(\frac{2\pi(t-1)}{n}\right) - 0.267 \cos\left(\frac{4\pi(t-1)}{n}\right) + 0.730 \sin\left(\frac{4\pi(t-1)}{n}\right) .$$

Based on Equation (29), the best-performing STSR–Fourier model for Banyuwangi employs two oscillation terms and an autoregressive coefficient of 0.686. This coefficient indicates that RH in the previous period has a positive influence on the current RH value. The inclusion of two oscillatory components suggests that the periodic structure of RH in Banyuwangi is more complex than that observed at the other stations, which require only one optimal oscillation term. Consequently, the model can effectively capture both the temporal dependence and seasonal fluctuations present in the Relative Humidity data.

Furthermore, graphical comparisons between the actual and predicted values for both the training and testing datasets are presented, together with the variation of the GCV values across different oscillation numbers considered in the STSR–Fourier model. In addition, diagnostic analyses of the fitted STSR–Fourier model are provided to assess model adequacy. These results are presented in Figures 11-14 and are used to evaluate the model's ability to capture the underlying data patterns, identify the optimal number of oscillations, and examine the validity of the model assumptions.

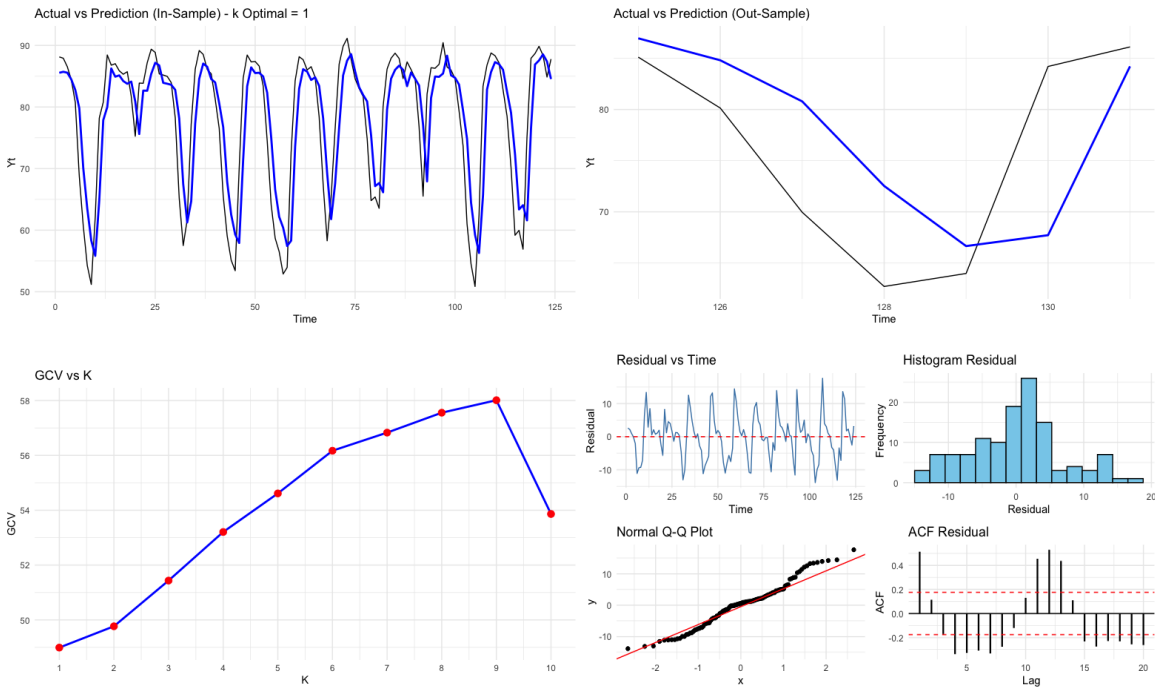


Figure 11. Evaluation Results of the STSR–Fourier Model at the Juanda

TIME-LAG SEMIPARAMETRIC MODEL OF RELATIVE HUMIDITY

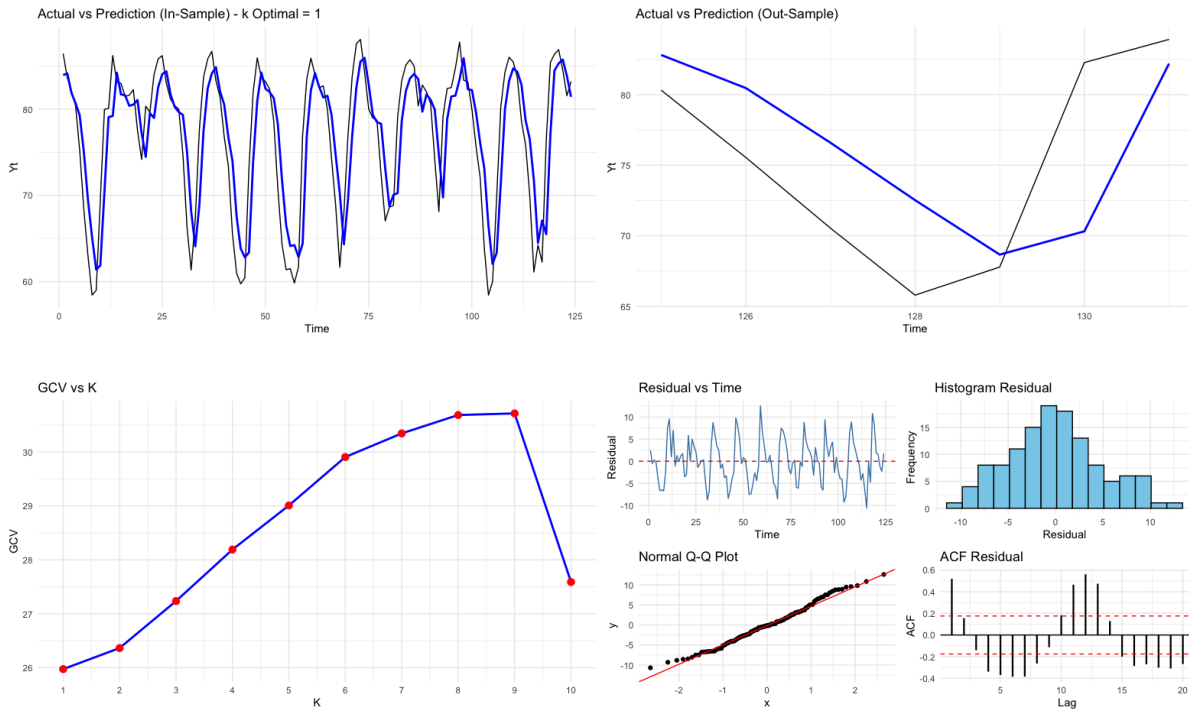


Figure 12. Evaluation Results of the STSR–Fourier Model at the Tanjung Perak

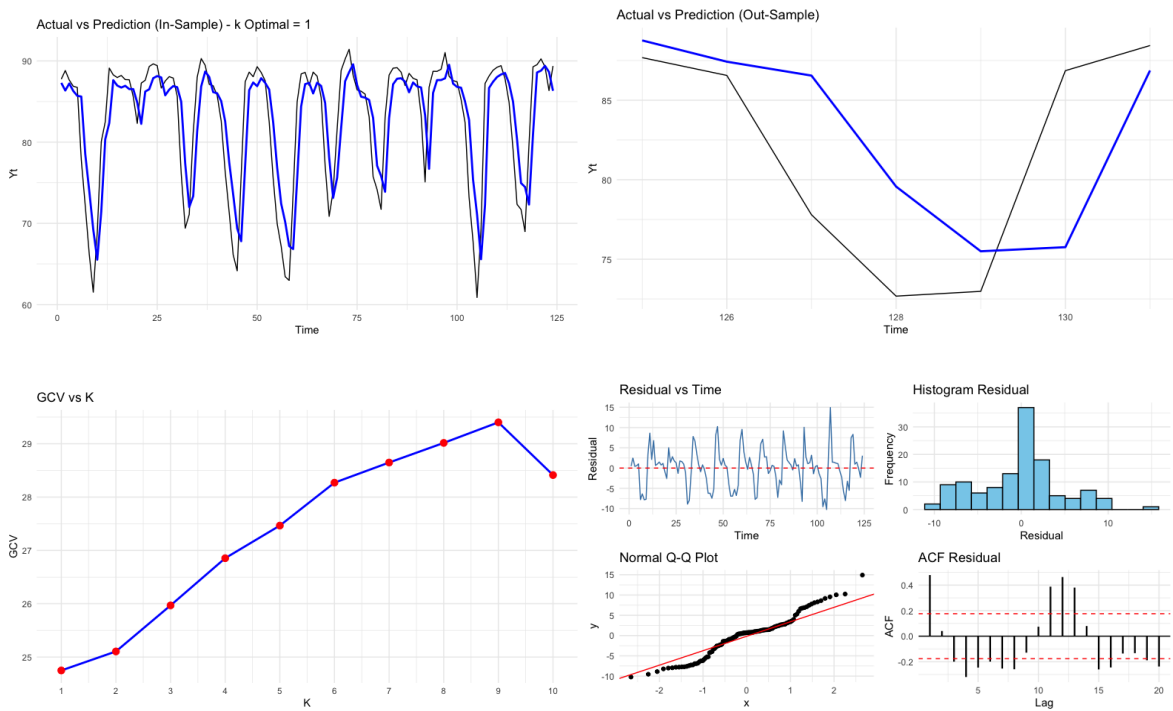


Figure 13. Evaluation Results of the STSR–Fourier Model at the Karangploso

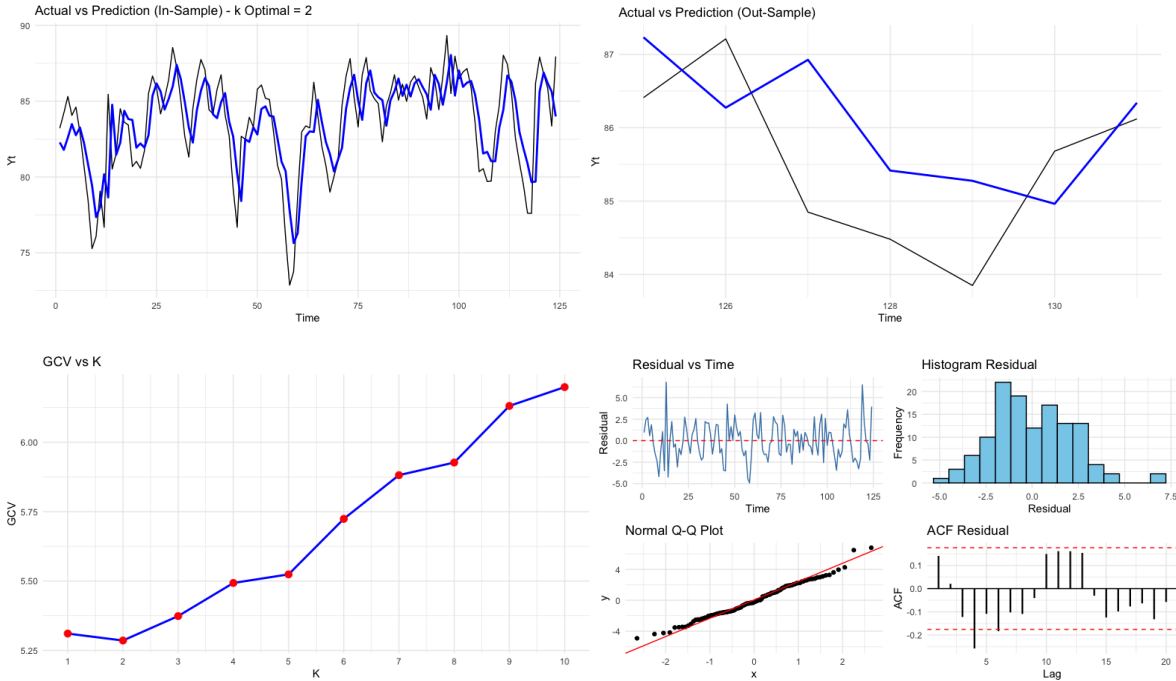


Figure 14. Evaluation Results of the STSR–Fourier Model at the Banyuwangi

Overall, as illustrated in Figures 11-14, the STSR-Fourier model demonstrates a strong ability to capture the underlying patterns of the Relative Humidity data in the training datasets. For the testing datasets, the model maintains good predictive performance, although slight deviations can be observed during several observation periods. Furthermore, the GCV value reaches its minimum at one oscillation for Juanda, Tanjung Perak, and Karangploso, and at two oscillations for Banyuwangi. Therefore, these oscillation numbers were selected as the optimal model parameters. Finally, the diagnostic results indicate that the model assumptions are reasonably satisfied, as evidenced by randomly distributed residuals, an approximately normal residual distribution, and the absence of significant autocorrelation.

4. CONCLUSION

This study investigated the modeling of Relative Humidity (RH) using FSNR, STSR-Fourier, and STSR-Local Polynomial based on MERRA-2 reanalysis data from four observation stations in East Java. The exploratory analysis revealed the presence of strong annual seasonal patterns and significant temporal dependence, indicating that modeling approaches capable of accommodating

both periodic behavior and time-lag effects are required to obtain a more representative model.

Based on the evaluation of three data partitioning schemes, the 95:5 training-testing scheme was selected as the optimal partition through a majority voting approach because it demonstrated the most consistent performance across the majority of models and observation stations. This finding suggests that a larger proportion of training data provides better estimation capability for capturing the seasonal and temporal characteristics of RH data. The comparative analysis showed that the STSR-Fourier model achieved the best predictive performance compared with the FSNR and STSR-Local Polynomial models. The integration of a first-order autoregressive component as the parametric part and Fourier Series as the nonparametric part was proven to improve prediction accuracy by producing lower testing MAPE values at most observation stations. These findings indicate that combining temporal dependence and periodic patterns within a semiparametric framework provides a better representation of RH dynamics than either a purely nonparametric approach or a local polynomial estimator.

From a methodological perspective, this study extends the application of Semiparametric Time Series Regression to satellite-based hydrometeorological reanalysis data and demonstrates that incorporating time-lag effects is an effective strategy for improving the accuracy of atmospheric data modeling characterized by seasonal and nonlinear patterns. Therefore, the STSR-Fourier model can be recommended as a reliable alternative for modeling Relative Humidity and other hydrometeorological variables. Future studies may consider the use of higher-order lags, the inclusion of additional meteorological variables, and the development of spatio-temporal semiparametric models to further improve predictive performance and provide a more comprehensive understanding of atmospheric phenomena.

FUNDING STATEMENT

This research was funded through a research grant provided by Universitas Brawijaya (Visiting Lecturer). The funding support was part of the university's internal research scheme aimed at encouraging academic research and innovation. The authors gratefully acknowledge this support, which made the completion of this study possible.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

- [1] J. Guo, L. Zhang, R. Guo, Relative Humidity Prediction with Covariates and Error Correction Based on SARIMA-EG-ECM Model, *Model. Earth Syst. Environ.* 9 (2023), 4493–4505. <https://doi.org/10.1007/s40808-023-01738-x>.
- [2] X. Wei, Y. Luo, X. Zhou, J. Zhao, H. Lu, et al., Temperature and Relative Humidity Prediction in South China Greenhouse Based on Machine Learning, *Sci. Rep.* 15 (2025), 24855. <https://doi.org/10.1038/s41598-025-08964-6>.
- [3] S. Zaharieva, I. Georgiev, S. Georgiev, A. Borodzhieva, V. Todorov, A Method for Forecasting Indoor Relative Humidity for Improving Comfort Conditions and Quality of Life, *Atmosphere* 16 (2025), 315. <https://doi.org/10.3390/atmos16030315>.
- [4] X. Xu, S. K. Frey, D. Ma, Hydrological Performance of ERA5 and MERRA-2 Precipitation Products over the Great Lakes Basin, *J. Hydrol. Reg. Stud.* 39 (2022), 100982. <https://doi.org/10.1016/j.ejrh.2021.100982>.
- [5] A. Wen, T. Wu, X. Wu, X. Zhu, R. Li, et al., Evaluation of MERRA-2 Land Surface Temperature Dataset and Its Application in Permafrost Mapping over China, *Atmos. Res.* 279 (2022), 106373. <https://doi.org/10.1016/j.atmosres.2022.106373>.
- [6] A. T. R. Dani, N. Chamidah, I. N. Budiantara, D. Aydin, Utilization of MERRA-2 Data for Precipitation Forecasting in Indonesia, Ibu Kota Nusantara: A Comparative Study of Traditional Methods and the Proposed Hybrid Fourier Series-ARIMA, in: 2025 International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2025, pp. 1–6. <https://doi.org/10.1109/IC3INA68387.2025.11325616>.
- [7] M. T. Nakkazi, J. I. Sempewo, M. D. Tumutungire, J. Byakatonda, Performance Evaluation of CFSR, MERRA-2 and TRMM3B42 Data Sets in Simulating River Discharge of Data-Scarce Tropical Catchments: A Case Study of Manafwa, Uganda, *J. Water Clim. Change* 13 (2022), 522–541. <https://doi.org/10.2166/wcc.2021.174>.
- [8] H. Tao, S. M. Awadh, S. Q. Salih, S. S. Shafik, Z. M. Yaseen, Integration of Extreme Gradient Boosting Feature Selection Approach with Machine Learning Models: Application of Weather Relative Humidity Prediction, *Neural Comput. Appl.* 34 (2022), 515–533. <https://doi.org/10.1007/s00521-021-06362-3>.
- [9] A. T. R. Dani, N. Chamidah, I. N. Budiantara, B. Lestari, D. Aydin, New Statistical Approach to Forecasting Earth's

- Skin Temperature from MERRA-2 Satellite Using Semiparametric Time Series Regression with Mixed Additive Spline Fourier (STSR-MASF), *Forecasting* 8 (2026), 6. <https://doi.org/10.3390/forecast8010006>.
- [10] Y. K. Truong, C. J. Stone, Semiparametric Time Series Regression, *J. Time Ser. Anal.* 15 (1994), 405–428. <https://doi.org/10.1111/j.1467-9892.1994.tb00202.x>.
- [11] J. Gao, Semiparametric Regression Smoothing of Non-Linear Time Series, *Scand. J. Stat.* 25 (1998), 521–539. <https://doi.org/10.1111/1467-9469.00118>.
- [12] J. Gao, H. Tong, Semiparametric Non-Linear Time Series Model Selection, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 66 (2004), 321–336. <https://doi.org/10.1111/j.1369-7412.2004.05303.x>.
- [13] J. Gao, M. King, Z. Lu, D. Tjøstheim, Nonparametric Specification Testing for Nonlinear Time Series with Nonstationarity, *Econom. Theory* 25 (2009), 1869–1892. <https://doi.org/10.1017/S0266466609990363>.
- [14] J. Chen, J. Gao, D. Li, Estimation in Semiparametric Time Series Regression, *Stat. Interface* 4 (2011), 243–251. <https://doi.org/10.4310/SII.2011.v4.n2.a18>.
- [15] J. Gao, K. Hawthorne, Semiparametric Estimation and Testing of the Trend of Temperature Series, *Econom. J.* 9 (2006), 332–355. <https://doi.org/10.1111/j.1368-423X.2006.00188.x>.
- [16] V. Fibriyani, N. Chamidah, T. Saifudin, Estimating Time Series Semiparametric Regression Model Using Local Polynomial Estimator for Predicting Inflation Rate in Indonesia, *J. King Saud Univ. Sci.* 36 (2024), 103549. <https://doi.org/10.1016/j.jksus.2024.103549>.
- [17] A. T. Fitriyah, N. Chamidah, T. Saifudin, Prediction of Paddy Production in Indonesia Using Semiparametric Time Series Regression Least Square Spline Estimator, *Data Metadata* 4 (2025), 527. <https://doi.org/10.56294/dm2025527>.
- [18] A. T. Fitriyah, N. Chamidah, T. Saifudin, B. Lestari, D. Aydin, Least Squares Spline Estimation Method in Semiparametric Time Series Regression for Predicting Indonesia Composite Index, *Stat. Optim. Inf. Comput.* 14 (2025), 3776–3803. <https://doi.org/10.19139/soic-2310-5070-2704>.
- [19] A. T. R. Dani, N. Chamidah, I. N. Budiantara, B. Lestari, R. Kusuma, et al., Semiparametric Time Series Regression Using Exponential Complex Fourier Series for Temperature Prediction in the Tropical Rainforest of Samarinda, *Innovaciencia* 13 (2025), e5364. <https://doi.org/10.15649/2346075x.5364>.
- [20] N. Chamidah, S. D. Febriana, R. A. Ariyanto, R. Sahawaly, Fourier Series Estimator for Predicting International Market Price of White Sugar, *AIP Conf. Proc.* 2329 (2021), 060035. <https://doi.org/10.1063/5.0042287>.

- [21] A. Ganesh, G. Balasubramanian, S. K. Jena, N. Pradhan, Fourier Approach to Function Approximation, *Int. J. Math. Arch.* 2 (2011), 617–624.
- [22] I. F. Amri, N. Chamidah, T. Saifudin, D. Purwanto, A. Fadlurohman, et al., Prediction of Extreme Weather Using Nonparametric Regression Approach with Fourier Series Estimators, *Data Metadata* 3 (2024), 319.
<https://doi.org/10.56294/dm2024319>.
- [23] R. Recylia, T. Saifudin, N. Chamidah, Forecasting the Number of Search and Rescue Operations for Ship Accidents in Indonesia Using Fourier Series Analysis (FSA), *BAREKENG: J. Math. & App.* 19 (2025), 1023–1036. <https://doi.org/10.30598/barekengvol19iss2pp1023-1036>.
- [24] J. Gao, *Nonlinear Time Series: Semiparametric and Nonparametric Methods*, Chapman and Hall/CRC, 2007.
<https://doi.org/10.1201/9781420011210>.
- [25] V. Fibriyani, N. Chamidah, Prediction of Inflation Rate in Indonesia Using Local Polynomial Estimator for Time Series Data, *J. Phys. Conf. Ser.* 1776 (2021), 012065. <https://doi.org/10.1088/1742-6596/1776/1/012065>.
- [26] V. Fibriyani, N. Chamidah, T. Saifudin, Modeling Case Fatality Rate of COVID-19 in Indonesia Using Time Series Semiparametric Regression Based on Local Polynomial Estimator, *Commun. Math. Biol. Neurosci.* 2024 (2024), 9. <https://doi.org/10.28919/cmbn/8379>.
- [27] H. Alabdulrazzaq, M. N. Alenezi, Y. Rawajfih, B. A. Alghannam, A. A. Al-Hassan, et al., On the Accuracy of ARIMA Based Prediction of COVID-19 Spread, *Results Phys.* 27 (2021), 104509.
<https://doi.org/10.1016/j.rinp.2021.104509>.
- [28] S. Makridakis, E. Spiliotis, V. Assimakopoulos, The Accuracy of Machine Learning (ML) Forecasting Methods versus Statistical Ones: Extending the Results of the M3-Competition, working paper, 2017.
<https://www.researchgate.net/publication/320298859>
- [29] C. Chen, J. Twycross, J. M. Garibaldi, A New Accuracy Measure Based on Bounded Relative Error for Time Series Forecasting, *PLoS ONE* 12 (2017), e0174202. <https://doi.org/10.1371/journal.pone.0174202>.
- [30] T. Juhana, H. Yuliana, Hendrawan, Iskandar, Y. Musashi, Comparative Analysis of ARIMA, Prophet, and Glmnet for Long Term Evolution (LTE) Base Station Traffic Forecasting, *Emerg. Sci. J.* 8 (2024), 2197–2217.
<https://doi.org/10.28991/ESJ-2024-08-06-04>.