

Article

An Expanded Spatial Durbin Model with Ordinary Kriging of Unobserved Big Climate Data

Annisa Nur Falah ¹, Yudhie Andriyana ^{2,*}, Budi Nurani Ruchjana ³, Eddy Hermawan ⁴, Teguh Harjana ⁴,
Edy Maryadi ⁵, Risyanto ⁴, Haries Satyawardhana ⁴ and Sinta Berliana Sipayung ⁴

¹ Post Doctoral Program, Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia; annisa15046@mail.unpad.ac.id

² Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia

³ Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia; budi.nurani@unpad.ac.id

⁴ Research Center for Climate and Atmosphere, National Research and Innovation Agency (BRIN), Jakarta Pusat 10340, Indonesia; eddy.hermawan@brin.go.id (E.H.); teguh.harjana@brin.go.id (T.H.); risyanto@brin.go.id (R.); hari031@brin.go.id (H.S.); sint001@brin.go.id (S.B.S.)

⁵ Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency (BRIN), Bandung 40135, Indonesia; edym004@brin.go.id

* Correspondence: y.andriyana@unpad.ac.id; Tel.: +62-82219223300

Abstract: Spatial models are essential in the prediction of climate phenomena because they can model the complex relationships between different locations. In this study, we discuss an expanded spatial Durbin model with ordinary kriging on unobserved locations (ESDMOK) to predict rainfall patterns in Java Island. The classical spatial Durbin model needed to be expanded to obtain a parameter estimation for each location. We combined this with ordinary kriging because the data were not available in some locations. The data were taken from the National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASA POWER) website. Since climate data are big data, we implement a big data analytics approach, namely the data analytics life cycle method. As the exogenous variables, we used air temperature, humidity, solar irradiation, wind speed, and surface pressure. The authors developed an R-Shiny web applications to implement our proposed technique. Using our proposed technique, we obtained more accurate and reliable climate data prediction, indicated by the mean absolute percentage error (MAPE), which was equal to 1.956%. The greatest effect on rainfall was given by the surface pressure variable, and the smallest was wind speed.

Keywords: expanded spatial Durbin model; ordinary kriging; data analytics life cycle; climate data

MSC: 6208; 62H11; 86A32



Citation: Falah, A.N.; Andriyana, Y.; Ruchjana, B.N.; Hermawan, E.; Harjana, T.; Maryadi, E.; Risyanto; Satyawardhana, H.; Sipayung, S.B. An Expanded Spatial Durbin Model with Ordinary Kriging of Unobserved Big Climate Data. *Mathematics* **2024**, *12*, 2447. <https://doi.org/10.3390/math12162447>

Academic Editor: MinJae Lee

Received: 30 June 2024

Revised: 26 July 2024

Accepted: 30 July 2024

Published: 7 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Climate is the long-term average weather and atmospheric conditions in a specific region, typically measured for 30 years or more, both regionally and globally [1]. These climate phenomena have significant effects on various aspects of human life, including agriculture, health, coastal ecosystems, and water quality [2]. One notable impact is La Nina, which increases rainfall in the Western Pacific region. Empirical data from BMKG indicates that La Nina can increase rainfall on Java Island by 20% and 70% [3]. Extreme rainfall causes landslides and floods, among other natural disasters [4,5]. The community suffers greatly from the financial and health effects of these natural disasters, and those who depend on rainfall conditions, like farmers and fishermen, truly need information related to rainfall prediction. The intensity of rainfall has a significant impact on every aspect of life, and exceptionally high rainfall has an impact on the occurrence of natural

disasters and so should be warned of before the event. People who depend on rainfall conditions, such as fishermen and farmers, really need information related to rainfall prediction. Rainfall is closely linked to Sustainable Development Goal (SDG) 13, which focuses on “Climate Action”. This goal emphasizes urgent action to combat climate change and its impacts [4]. Understanding and managing rainfall patterns are crucial for achieving SDG 13. Effective climate action requires addressing the impacts of changing rainfall on various sectors, enhancing resilience to extreme weather events, and ensuring sustainable water and land management.

The spatial model is an effective tool for predicting rainfall and climate. Creating statistical or computational models to represent spatial relationships or patterns in data is a prevalent practice in spatial modelling [5]. Spatial models are widely used in various fields to analyse patterns, relationships, and phenomena related to the spatial dimension in data [6]. Rainfall is closely related to other climate elements, such as air temperature, humidity, solar irradiation, wind speed and surface pressure, which vary from region to region [7]. These climate variables are observed and modelled using spatial analysis. Falah et al. (2023) conducted hybrid modelling of the Spatial Autoregressive Exogenous (SAR-X) model using Casetti’s model approach for the prediction of rainfall in West Java, Indonesia [8]. The weakness of the SAR-X model is that it only facilitates spatial dependence on the response variable, not paying attention to spatial dependence on the exogenous variables. Discussing spatial aspects, there is a possibility that there is spatial dependency not only on the response variable but also on exogenous variables. Thus, a model approach is needed that can facilitate the existence of spatial dependencies between these variables. Therefore, to overcome these problems, the expanded spatial Durbin model (ESDM) was used to overcome spatial dependence on exogenous variables [9].

In various phenomena, there are often unknown values at observation locations. A linear interpolation model called the ordinary kriging (OK) model is utilized to predict these values at unobserved locations [10]. The OK method is a spatial interpolation method that uses the spatial variability of data to estimate values at unobserved locations [11]. The advantage of the OK method is that it produces an optimal estimate by taking into account the spatial information and covariance structure of the data [12]. This is to obtain the most accurate prediction at unobserved locations [13]. The predictions generated by OK often produce smooth maps without any sharp spikes or dips between observation points [14].

The ongoing expansion of data poses a challenge to the prediction of climate and rainfall patterns. We discovered a gap in rainfall prediction utilizing the ESDM with OK of climate data based on the summary above. These climate data were collected from 119 districts/cities in Java Island using the National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASA POWER) website. In this research, the data analytics life cycle was the big data approach methodology, which was utilized to predict rainfall using big climate data. Rainfall description and prediction have a more in-depth stage in the data analytics life cycle. The application study in this research is supported by computing an integrated R script for the development of the ESDM using an R-Shiny web application to facilitate the prediction process.

2. Materials and Methods

2.1. Experimental Semivariogram

The OK method was used to analyse geostatistical data and interpolate values based on observed data. It is worth noting that D. G. Krige, a gold mining expert from South Africa, first introduced the model. Furthermore, OK relies on the assumption that there is a spatial correlation between observed data as determined by the distance between the entities [15]. To perform the interpolation, the model employs a semivariogram calculation representing the spatial differences and values between all pairs of observed data. The semivariogram also shows the weights used in the interpolation process, which is calculated

based on a sample semivariogram with distance h , different z values, and n data samples. The experimental semivariogram at distance h can be expressed as follows [16]:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(s_i + h) - Z(s_i)]^2 \tag{1}$$

where:

- $\hat{\gamma}(h)$: experimental semivariogram value with distance h ;
- $Z(s_i)$: observation value in location s_i ;
- $Z(s_i + h)$: observation value in location $s_i + h$;
- $N(h)$: many pairs of data with the same distance h ;
- h : distance.

All possible pairs of distances were calculated using a Euclidean distance equation as follows:

$$|h| = \sqrt{(s_i(u_i) - s_j(u_j))^2 - (s_i(v_i) - s_j(v_j))^2} \tag{2}$$

The results of the Euclidean distance calculation were converted to kilometres with $|d_{ij}| \times 111.319$. A value of 111.319 was obtained by converting 1 degree longitude to kilometres [17]. A location can be determined as $s_{ij}(u_{ij}, v_{ij})$, where s_{ij} is the symbol of the location i and j , with $i = 1, 2, 3 \dots, N$ and $j = 1, 2, 3 \dots, N$, while u and v indicate the latitude and longitude coordinates.

2.2. Theoretical Semivariogram Model

There are three commonly used theoretical semivariogram models in kriging: the spherical, the Gaussian, and the exponential [16]. Furthermore, theoretical semivariogram values can be divided into three using the equations in Table 1: h is the location distance between samples, c is the sill value, and a is the range [18]. Theoretical semivariogram models can be seen in Table 1, and a theoretical semivariogram plot can be seen in Figure 1.

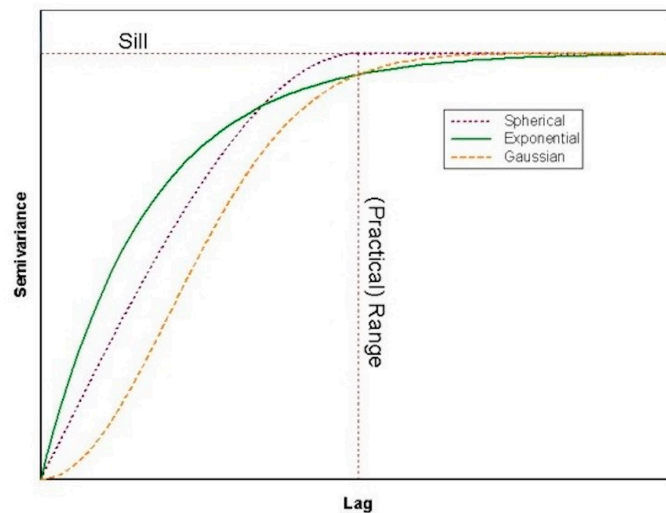


Figure 1. Theoretical semivariogram model plot.

Table 1. Theoretical semivariogram model.

Model	Function
Spherical	$\gamma(h) = \begin{cases} c \left[\left(\frac{3h}{2a} \right) - \left(\frac{h}{2a} \right)^3 \right], & h \leq a \\ c, & h > a \end{cases} \quad (3)$
Exponential	$\gamma(h) = \begin{cases} c \left[1 - \exp\left(-\frac{h}{a}\right) \right], & h \leq a \\ c, & h > a \end{cases} \quad (4)$
Gaussian	$\gamma(h) = \begin{cases} c \left[1 - \exp\left(-\frac{h}{a}\right)^2 \right], & h \leq a \\ c, & h > a \end{cases} \quad (5)$

2.3. Ordinary Kriging (OK) Method

The kriging method is a prediction method that provides a BLUE (best linear unbiased estimation) estimator of point values or averages for prediction at unobserved locations. This method uses a semivariogram calculation that represents the spatial and value differences between all pairs of data samples. According to [19], the kriging estimator $\hat{Z}(s)$, where s is the location at the unsampled point, is a linear combination of random variables; this can be seen in the kriging estimation formulated in the following equation:

$$\hat{Z}(s) - m(s) = \sum_{i=1}^n \lambda_i [Z(s_i) - m(s_i)] \quad (6)$$

with:

- s : predicted locations;
- s_i : i -th data location adjacent to the predicted location;
- $m(s)$: the expected or average value of $Z(s)$;
- $m(s_i)$: the expected or average value of $Z(s_i)$;
- n : the number of data used for prediction;
- λ_i : weight value at i -th location.

The objective of the kriging method is to determine the weight values λ_i that result in minimum estimator variance and an unbiased estimator. The estimator variance can be expressed as follows [20]:

$$\sigma_e^2(s) = Var[\hat{Z}(s) - Z(s)] \quad (7)$$

while the requirement to produce an unbiased estimator is:

$$E[\hat{Z}(s) - Z(s)] = 0 \quad (8)$$

The OK method is one of the kriging methods that assumes that the mean is unknown. If $Z(s) = m(s)$ and $Z(s_i) = m(s_i)$, then $m(s) = m(s_i) = m$, and then Equation (6) becomes:

$$\hat{Z}(s) - m = \sum_{i=1}^n \lambda_i [Z(s_i) - m]$$

$$\hat{Z}(s) = m + \sum_{i=1}^n \lambda_i [Z(s_i) - m]$$

$$\hat{Z}(s) = \sum_{i=1}^n \lambda_i Z(s_i) - m \left(\sum_{i=1}^n \lambda_i - 1 \right)$$

Since m is assumed to be unknown, the OK estimator is obtained as $\hat{Z}(s) = \sum_{i=1}^n \lambda_i Z(s_i)$

with the condition $\sum_{i=1}^n \lambda_i = 1$.

The BLUE (best linear unbiased estimation) properties of the OK method are shown as follows:

- Linear

The OK estimator obtained from n observations of the data used forms a linear model, namely:

$$\hat{Z}(s) = \sum_{i=1}^n \lambda_i Z(s_i) \tag{9}$$

- Unbiased

The OK estimator is unbiased if it satisfies Equation (8)

$$\begin{aligned} E[\hat{Z}(s_i) - Z(s)] &= E\left[\sum_{i=1}^n \lambda_i Z(s_i) - Z(s)\right] \\ &= \sum_{i=1}^n \lambda_i E[Z(s_i) - Z(s)] \quad (\because E[Z(s_i) - Z(s)] = 0) \\ &= 0 \end{aligned}$$

since the mean is assumed to be unknown, $E[Z(s_i) - Z(s)] = 0$, and the unbiased estimator property is satisfied by the OK method.

- Best

The best here means that the OK estimator has the minimum error variance. The variance of the estimator of OK is as follows:

$$\begin{aligned} \sigma_{OK}^2 &= Var[\hat{Z}(s) - Z(s)] \\ \sigma_{OK}^2 &= Var[\hat{Z}(s)] + Var[Z(s)] - 2Cov[\hat{Z}(s), Z(s)] \end{aligned} \tag{10}$$

to describe $Var[\hat{Z}(s)]$ in Equation (10), it is known that $\hat{Z}(s) = \sum_{i=1}^n \lambda_i Z(s_i)$, which can be expressed as follows:

$$\begin{aligned} Var[\hat{Z}(s)] &= Var\left[\sum_{i=1}^n \lambda_i Z(s_i)\right] \\ Var[\hat{Z}(s)] &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j Cov[Z(s_i), Z(s_j)] \end{aligned} \tag{11}$$

to describe $Cov[\hat{Z}(s), Z(s)]$ in Equation (11):

$$\begin{aligned} Cov[\hat{Z}(s), Z(s)] &= E[\hat{Z}(s)Z(s)] - E[\hat{Z}(s)]E[Z(s)] \\ Cov[\hat{Z}(s), Z(s)] &= E\left[\left(\sum_{i=1}^n \lambda_i Z(s_i)\right)Z(s)\right] - E\left[\sum_{i=1}^n \lambda_i Z(s_i)\right]E[Z(s)] \\ Cov[\hat{Z}(s), Z(s)] &= \sum_{i=1}^n \lambda_i E[Z(s_i)Z(s)] - \sum_{i=1}^n \lambda_i E[Z(s_i)]E[Z(s)] \\ Cov[\hat{Z}(s), Z(s)] &= \sum_{i=1}^n \lambda_i Cov[Z(s_i), Z(s)] \end{aligned} \tag{12}$$

Supposing that $Var[Z(s)] = \sigma^2$, Equations (11) and (12) can be substituted into Equation (10), and the following equation can be obtained:

$$\sigma_{OK}^2 = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j Cov[Z(s_i), Z(s_j)] + \sigma^2 - 2 \sum_{i=1}^n \lambda_i Cov[Z(s_i), Z(s)] \tag{13}$$

with the condition $\sum_{i=1}^n \lambda_i = 1$.

Based on Equation (13), to obtain the minimum value of the estimator variance, the Lagrange multiplier (LM) method was used with the parameter μ . The LM equation is expressed as follows:

$$F(\lambda_i, \mu) = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \text{Cov}[Z(s_i), Z(s_j)] + \sigma^2 - 2 \sum_{i=1}^n \lambda_i \text{Cov}[Z(s_i), Z(s)] + 2\mu \left[\sum_{i=1}^n \lambda_i - 1 \right] \tag{14}$$

By deriving Equation (14) for the following variables λ_i :

$$\frac{\partial F(\lambda_i, \mu)}{\partial \lambda_i} = 2 \sum_{j=1}^n \lambda_j \text{Cov}[Z(s_i), Z(s_j)] - 2 \text{Cov}[Z(s_i), Z(s)] + 2\mu = 0$$

Because $\frac{\partial F(\lambda_i, \mu)}{\partial \lambda_i} = 0$, the following is obtained:

$$\sum_{j=1}^n \lambda_j \text{Cov}[Z(s_i), Z(s_j)] = \text{Cov}[Z(s_i), Z(s)] - \mu \tag{15}$$

By deriving Equation (14) with respect to the parameter μ , we obtain:

$$\frac{\partial F(\lambda_i, \mu)}{\partial \mu} = 2 \left[\sum_{i=1}^n \lambda_i - 1 \right] = 0$$

Because $\frac{\partial F(\lambda_i, \mu)}{\partial \mu} = 0$, the following is obtained:

$$\begin{aligned} 2 \left[\sum_{i=1}^n \lambda_i - 1 \right] &= 0 \\ \sum_{i=1}^n \lambda_i - 1 &= 0 \end{aligned}$$

$$\sum_{i=1}^n \lambda_i = 1 \tag{16}$$

Equations (15) and (16) compose an OK system. By solving the equation, the following matrix can be formed:

$$\begin{aligned} \lambda_1 C_{11} + \lambda_2 C_{12} + \lambda_3 C_{13} + \dots + \lambda_n C_{1n} + \mu &= C_{10} \\ \lambda_1 C_{21} + \lambda_2 C_{22} + \lambda_3 C_{23} + \dots + \lambda_n C_{2n} + \mu &= C_{20} \\ \lambda_1 C_{31} + \lambda_2 C_{32} + \lambda_3 C_{33} + \dots + \lambda_n C_{3n} + \mu &= C_{30} \\ &\vdots \\ \lambda_1 C_{n1} + \lambda_2 C_{n2} + \lambda_3 C_{n3} + \dots + \lambda_n C_{nn} + \mu &= C_{n0} \\ \lambda_1 + \lambda_2 + \lambda_3 + \dots + \lambda_n + 0 &= 1 \end{aligned}$$

$$\begin{pmatrix} C_{11} & C_{12} & C_{13} & \dots & C_{1n} & 1 \\ C_{21} & C_{22} & C_{23} & \dots & C_{2n} & 1 \\ C_{31} & C_{32} & C_{33} & \dots & C_{3n} & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{n1} & C_{n2} & C_{n3} & \dots & C_{nn} & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} C_{10} \\ C_{20} \\ C_{30} \\ \vdots \\ C_{n0} \\ 1 \end{pmatrix} \tag{17}$$

Meanwhile, to determine the weight value of each observed point against the unobserved point, this can be expressed as follows:

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & C_{13} & \cdots & C_{1n} & 1 \\ C_{21} & C_{22} & C_{23} & \cdots & C_{2n} & 1 \\ C_{31} & C_{32} & C_{33} & \cdots & C_{3n} & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{n1} & C_{n2} & C_{n3} & \cdots & C_{nn} & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} C_{10} \\ C_{20} \\ C_{30} \\ \vdots \\ C_{n0} \\ 1 \end{pmatrix} \tag{18}$$

$$\lambda = C_{nn}^{-1} C_{n0}$$

where:

C_{nn} : the variance covariance matrix between the variables sampled at the location n and the sampled variable at location n ;

C_{n0} : the variance vector of the covariance between the sampled variable at the location n and the predicted variable;

μ : the Lagrange multiplier parameter.

To obtain the variance equation of the OK estimator in Equation (13), Equation (15) can be substituted, and the following is obtained:

$$\begin{aligned} \sigma_{OK}^2 &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j Cov[Z(s_i), Z(s_j)] + \sigma^2 - 2 \sum_{i=1}^n \lambda_i Cov[Z(s_i), Z(s)] \\ \sigma_{OK}^2 &= \sum_{i=1}^n \lambda_i \sum_{j=1}^n \lambda_j Cov[Z(s_i), Z(s_j)] + \sigma^2 - 2 \sum_{i=1}^n \lambda_i Cov[Z(s_i), Z(s)] \\ \sigma_{OK}^2 &= \sum_{i=1}^n \lambda_i Cov[Z(s_i), Z(s)] - \mu + \sigma^2 - 2 \sum_{i=1}^n \lambda_i Cov[Z(s_i), Z(s)] \end{aligned}$$

$$\sigma_e^2 = \sigma^2 - \sum_{i=1}^n \lambda_i Cov[Z(s_i), Z(s)] - \mu$$

$$\sigma_{OK}^2 = \sigma^2 - (\lambda_1 C_{10} + \lambda_2 C_{20} + \lambda_3 C_{30} + \dots + \lambda_n C_{n0}) - \mu$$

$$\sigma_{OK}^2 = \sigma^2 - (C_{10} \ C_{20} \ C_{30} \ \cdots \ C_{n0} \ 1) \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} \tag{19}$$

The minimum estimator variance is commonly referred to as the OK estimator variance; thus, the best estimator satisfied the OK method.

2.4. Expanded Spatial Durbin Model (ESDM)

The ESDM was used to overcome spatial dependence on the exogenous variables. The ESDM is formulated as follows [9]:

$$y = \rho W y + \alpha \mathbf{1}_n + XZJ\beta_0 + W\tilde{X}\theta + \epsilon \text{ with } \epsilon \stackrel{iid}{\sim} N(0, \sigma^2 \mathbf{I}) \tag{20}$$

Letting $A = XZJ$, it follows that:

$$y = \rho W y + \alpha \mathbf{1}_n + A\beta_0 + W\tilde{X}\theta + \epsilon \text{ with } \epsilon \stackrel{iid}{\sim} N(0, \sigma^2 \mathbf{I}) \tag{21}$$

$$y = \rho W y + U\delta + \epsilon \tag{22}$$

where:

$$U = [1_n \quad A \quad W\tilde{X}],$$

$$\delta = \begin{bmatrix} \alpha \\ \beta_0 \\ \theta \end{bmatrix},$$

with:

- y**: vector of dependent variables of size $(n \times 1)$;
- \tilde{X} : matrix of independent variables of size $(n \times k)$;
- X**: matrix of independent variables of size $(n \times nk)$;
- ρ : spatial lag coefficient of the dependent variable;
- α : constant parameter;
- W**: spatial weight matrix of size $(n \times n)$;
- Z**: location information that contains elements Z_{xi}, Z_{yi} with $i = 1, \dots, n$, representing the latitude and longitude of each observation, of size $(nk \times 2nk)$;
- J**: expansion of the identity matrix of size $(2nk \times 2k)$;
- β : matrix of size $(nk \times 1)$, which contains parameter estimators for all explanatory k variables at each observation;
- β_0 : parameter expressed by $\beta_{latitude}, \beta_{longitude}$ of size $(2k \times 1)$;
- θ : spatial lag parameter vector of covariate variable of size $(k \times 1)$;
- \otimes : Kronecker product;
- ϵ : error vector of size $(n \times 1)$;
- s_i : location matrix with $i = 1, \dots, n$.

2.5. Mean Absolute Percentage Error (MAPE)

To evaluate the model’s performance, the mean absolute percentage error (MAPE) is calculated as follows:

$$MAPE = \left(\frac{1}{n} \sum_{i=1}^n \left| \frac{y(s_i) - \hat{y}(s_i)}{y(s_i)} \right| \right) \times 100\% \tag{23}$$

with

- $y(s_i)$: the values in the actual data at the location s_i ;
- $\hat{y}(s_i)$: the values in the prediction data at the location s_i ;
- n : the number of observation locations.

According to Lawrence’s criteria (2009) [21], MAPE values can be categorized as follows (Table 2):

Table 2. MAPE score scale.

Scale MAPE	Accuracy Score
$\leq 10\%$	Very accurate prediction
$10 < MAPE \leq 20\%$	Good prediction
$20 < MAPE \leq 50\%$	Reasonable prediction
$> 50\%$	Inaccurate prediction

2.6. Data Analytics Life Cycle

Large data quantities, a variety of data architectures, and rapid growth in data were the challenges faced by big data and data science, which led to the creation of the data analytics life cycle. There are six stages in this life cycle, which might happen simultaneously in some circumstances. This analysis could typically proceed both forward and backward, enabling an iterative process that takes into account newly discovered information as it becomes available [22]. This makes it possible to solve problems and go through the procedure

repeatedly, which also makes it easier to operationalize research objectives. Best practices for the analytical process, from discovery to research work completion, are defined by the data analysis life cycle. The data analytics life cycle consists of six stages that apply looping or backward/forward. The following is a summary of the six stages of the data analysis life cycle [23]:

- **Discovery (Problem Formulation):** At this point, a literature review was conducted to prepare for the research problem analysis phase. This phase required assembling resources, including data, technology, references, and time. Developing a problem framework as an analytical task to be tackled in the following stage and developing preliminary hypotheses to investigate and evaluate the data were crucial tasks in this phase.
- **Data Preparation:** Initial data analysis was part of the data pre-processing performed at this stage. A necessary step before building the model was to prepare the data for collection in the database repository, which included procedures such as data cleaning, extraction, transformation, and integration.
- **Model Planning:** This stage focuses on planning the model by determining the methods, techniques, and research flow to be followed during the model-building stage.
- **Model Building:** At this stage, this research focused on creating datasets for testing, training, and creating output models. The model's efficiency in running on the current hardware, such as its quick hardware and parallel processing capabilities, was considered.
- **Communicating Results:** This step entailed testing the data model and any modifications with the user or in an experimental environment to ascertain whether the output complied with the development criteria. Should the model fail to satisfy the specifications, an assessment was carried out, and the procedure could revert to the earlier phase for further improvement.
- **Operationalizing (Operationalization):** At this point, the final report, instructions, codes, and technical documents had to be submitted. To guarantee a wider application, this stage can also include implementing the model as a pilot project.

If more enhancements are needed, the data analytics life cycle might be carried out again from phases 1 through 5. The evaluation of the modelling process, from steps 6 to 1, was indicated by dotted lines, highlighting the possibility of revisiting certain stages if the modelling results did not meet the desired criteria.

3. Results

3.1. Data Description

The objective of this study was to predict big climate data using the ESDMOK in Java Island, Indonesia. The model was applied to secondary data obtained from the National Aeronautics and Space Administration (NASA) Prediction of Worldwide Energy Resources (POWER). The POWER project provides solar and meteorological data generated by NASA to support renewable energy, building energy efficiency, and agricultural needs. The POWER project started in 2003 as an outgrowth of Surface meteorology and Solar Energy (SSE). NASA-generated satellite data are essential in supporting researchers and the public in studying Earth's climate and climate processes [24]. The POWER project provides long-term climatological mean estimates of meteorological data and solar energy flux surface data. In addition to these long-term climatological averages, daily data in the form of time series are also available. Solar data are based on satellite observations, and meteorological data are derived from the MERRA-2 assimilation model. These satellite and model-based products have proven to be quite accurate in providing reliable solar and meteorological resource data in regions where surface measurements could be more sparse or non-existent. The uncertainty estimates of POWER data are based on comparisons with measurement data [25–27].

In addition, POWER also provides high-resolution precipitation data derived from NASA's Global Precipitation Measurement (GPM) mission's Integrated Multi-satellite

Retrievals for GPM (IMERG) with a resolution of $0.5^\circ \times 0.625^\circ$ latitude–longitude grid (approximately 50 km) [28]. NASA POWER data can be downloaded free via the web at <https://power.larc.nasa.gov/> (accessed on 5 July 2023). The data retrieval process began with inputting the latitude and longitude coordinates of the location and determining the time interval. In this study, the daily data interval was taken from 1 January 1982 to 5 July 2023 and was recorded in a daily data format. The selected climate variables included rainfall, air temperature, humidity, wind speed, solar irradiation, and surface pressure, along with latitude and longitude coordinate information. The output data obtained were stored in files of comma-separated value (.csv) format.

3.2. Data Analytics Life Cycle for the ESDMOK

In this research, rainfall prediction with ESDMOK uses the data analytics life cycle methodology shown in Figure 2. The process begins with formulating research problems, including natural disasters caused by rainfall, problem identification, climate variables affecting rainfall levels, and initial hypotheses based on theories that support ESDMOK. Furthermore, the data preparation stage includes determining the source of climate data to be analyzed and data pre-processing. The data collection process begins with inputting information on the location coordinates of 119 districts/cities on the Java Island, determining the observation time interval in the form of daily data, and selecting climate variables. In the data collection process, this research utilizes the application programming interface (API) by running the “pynasapower” package in Python. The data pre-processing stage includes; removing missing values, aggregating daily data to monthly data, selecting the climate variables. The model planning stage integrates location data with climate variables to provide input data for ESDMOK, such as spatial modelling. Furthermore, at the model development stage, based on the NASA POWER grid resolution of latitude-longitude $0.5^\circ \times 0.625^\circ$ (approximately 50 km), 55 locations with the same data were generated. Thus, the locations were split into 64 observed locations and 55 unobserved locations, and the prediction stage was performed using ESDMOK. The communication results stage includes the model evaluation stage, which calculates accuracy using the MAPE calculation, the post-processing stage, which visualizes spatial mapping for rainfall prediction, and the interpretation of results to gain knowledge that can be used as a recommendation. The last stage is operationalization, which involves documenting research results and disseminating scientific papers in journals.

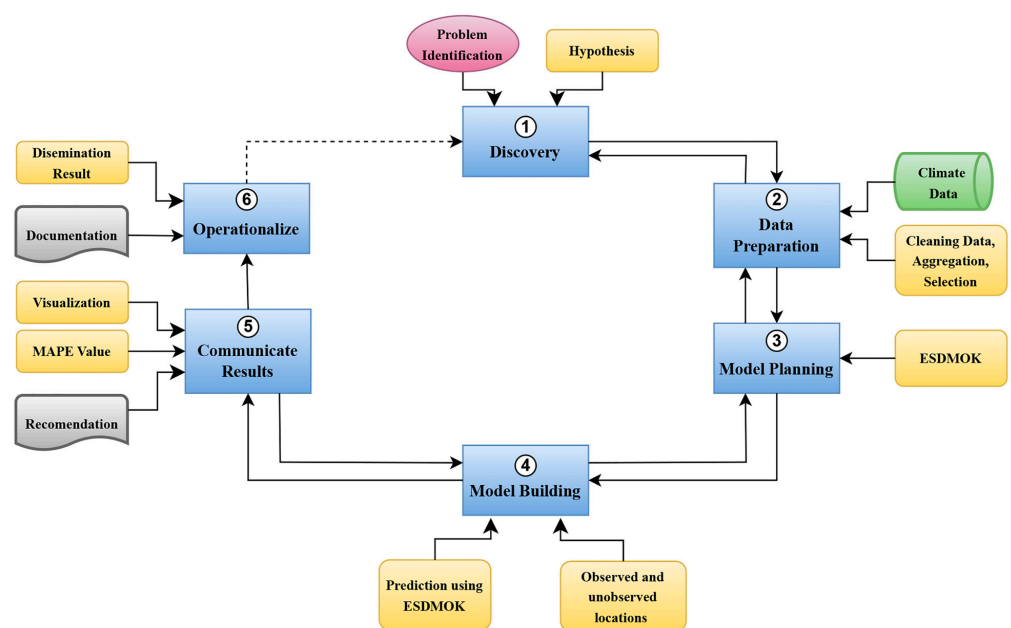


Figure 2. Data analytics life cycle methodology of the ESDMOK.

3.3. Framework ESDMOK for Prediction

Based on the gap analysis, ESDMOK is a spatial model that can predict unobserved locations and considers spatial dependencies in exogenous variables. Figure 3 shows the framework diagram for ESDMOK planning as part of the model-building stage. As mentioned in the data preparation stage, the process starts with inputting climate data sourced from NASA POWER. Next, in the data pre-processing stage, removing missing values (−999), aggregating daily data into monthly data, and selecting duplicate data. In this study, climate data from NASA POWER went through a pre-processing data stage using an R-Shiny web application, available at the following link: <https://annisanurfalah.shinyapps.io/Pre-ProcessingData/> (accessed on 26 September 2023). The resulting pre-processed data were split into observed and unobserved locations, which can be seen in Appendix A. The observed locations were used as an input in the OK method for predicting climate data at unobserved locations. In this study, predictions at unobserved locations were calculated with an OK method using an R-Shiny web application, available at the following link: <https://annisanurfalah.shinyapps.io/Ordinary-Point-Kriging/> (accessed on 7 May 2024). The integration results of climate data at observed and unobserved locations were used to construct an inverse distance weight matrix and spatial autocorrelation was assessed using the Moran Index and Scatterplot. If spatial autocorrelation is detected, the process continues with ESDM, using the Maximum Likelihood Estimation (MLE) method. Parameter estimates for ESDM are then calculated, followed by evaluating prediction accuracy using Mean Absolute Percentage Error (MAPE). The prediction calculations used the ESDM via an R-Shiny web application, available at the following link: <https://andriyanafalah.shinyapps.io/SDM-Expansion/> (accessed on 11 June 2024). The prediction results are then processed by visualizing the spatial mapping in the form of web application-based maps, choropleth maps and providing interpretation to gain valuable insights.

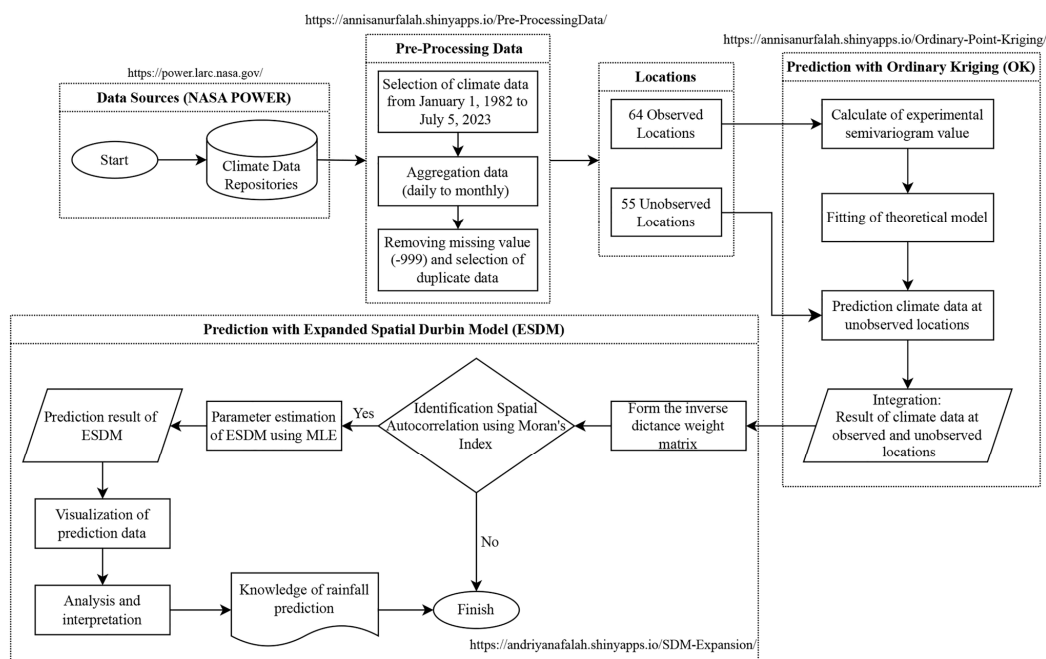


Figure 3. Framework diagram of ESDMOK for prediction.

3.4. Prediction Result of OK Method at Unobserved Locations

Semivariogram values were calculated based on all possible distance pairs, where the distance function used was the Euclidean distance, a function of the distance h , which describes the difference between the main variable and the difference in the additional variable h . Equation (1) was used to determine the semivariogram value and the number

of distance pairs. The research data that were observed consisted of 64 districts/cities for six climate data categories, including rainfall, air temperature, humidity, wind speed, solar irradiation, and surface pressure. The calculation results of the experimental semivariogram values are shown in Table 3.

Table 3. Experimental semivariogram values.

No.	The Number of Data Points with the Same Distance	Distance	Experimental Semivariogram Values					
			Rainfall	Air Temperature	Humidity	Wind Speed	Solar Irradiation	Surface Pressure
1	9	16,986.956	12.612	0.865	1.904	0.186	0.033	2.442
2	40	32,999.848	68.209	1.389	4.282	0.358	0.074	3.495
3	76	54,226.021	72.565	1.429	5.398	0.232	0.107	3.252
4	99	74,327.202	90.908	1.789	5.820	0.285	0.144	4.465
5	96	96,082.271	168.067	2.390	8.202	0.400	0.133	5.657
6	109	117,949.815	159.668	2.880	9.046	0.327	0.180	7.131
7	94	138,289.058	169.586	2.650	8.612	0.386	0.204	5.904
8	90	159,416.873	219.595	2.116	9.571	0.299	0.295	4.560
9	86	181,276.821	265.071	3.191	10.495	0.388	0.297	7.804
10	76	202,391.4	335.187	2.184	9.989	0.399	0.264	4.687
11	83	223,437.78	377.226	2.491	12.527	0.346	0.340	5.015
12	72	244,803.453	371.607	2.246	11.461	0.280	0.419	4.565
13	84	268,191.079	510.885	1.918	12.129	0.429	0.482	3.515
14	55	287,619.435	534.357	3.059	18.205	0.300	0.424	5.342
15	63	309,682.512	581.690	2.848	14.276	0.359	0.488	6.496

The experimental semivariogram is used for fitting the theoretical semivariogram model. Reference to Table 1, the theoretical semivariogram model with varying sill and range values in each climate data set was fitted to the experimental semivariogram values. The theoretical semivariogram was fit for climate data, and the results are plotted in Appendix B. The sum square error (SSE) number was used to determine which theoretical semivariogram model was the best. The SSE values for the spherical, exponential, and Gaussian models are shown in Table 4. Based on the lowest SSE values, indicated by bold numbers, the optimal models for the rainfall, air temperature, humidity, solar irradiation, wind speed, and surface pressure are chosen.

Table 4. Theoretical semivariogram for OK method.

SSE	Spherical	Exponential	Gaussian
Rainfall	0.0001093	0.0001135	8.86 × 10⁻⁵
Air Temperature	6.96 × 10 ⁻⁹	8.56 × 10 ⁻⁹	8.93 × 10⁻⁹
Humidity	1.35 × 10 ⁻⁷	8.59 × 10⁻⁸	6.28 × 10 ⁻⁸
Wind Speed	1.92 × 10 ⁻⁸	5.35 × 10⁻¹⁰	4.47 × 10 ⁻¹⁰
Solar Irradiation	6.76 × 10 ⁻¹⁰	8.82 × 10⁻¹¹	4.27 × 10 ⁻¹¹
Surface Pressure	7.11 × 10 ⁻⁸	7.82 × 10⁻⁸	7.33 × 10 ⁻⁸

Based on Table 4, the Gaussian semivariogram model had the minimum SSE value for rainfall and air temperature. Then, the exponential semivariogram model had the minimum SSE value for humidity, wind speed, solar irradiation and surface pressure. Therefore, these models were selected as the input for the calculation prediction of the OK method at unobserved locations (55 districts/cities). The integration results at observed locations (64 districts/cities) and unobserved locations are presented in the form of spatial mapping visualization, which can be seen in Figure 4. Climate variables in this data visualization include rainfall, air temperature, humidity, wind speed, solar irradiation, and surface pressure, which are variables in 119 districts/cities in Java Island, Indonesia. The data of these climate variables are represented in colours and bars, namely in several colours or one colour with different levels of intensity, to provide information on the level of value of the data. This data visualization aims to convey the value of climate variables so that data visualization can be classified into the explanation category. Still, with the addition

of several interactive features and for easy use, this data visualization was made in the form of a web application. Data visualization in the form of spatial mapping based on web applications was developed using the Javascript programming language, HyperText Markup Language (HTML), and Cascade Style Sheets (CSS) style sheet language. Javascript was used to modify the display to handle data processing and data structures, while HTML and CSS were used for the content of the webpage, namely the layout and structure of the display. Several libraries were used in this data visualization, namely the leaflet library as a base map, map features such as overlay, zoom in, zoom out, and pan, and another library including highchart for data representation in the form of bar charts.

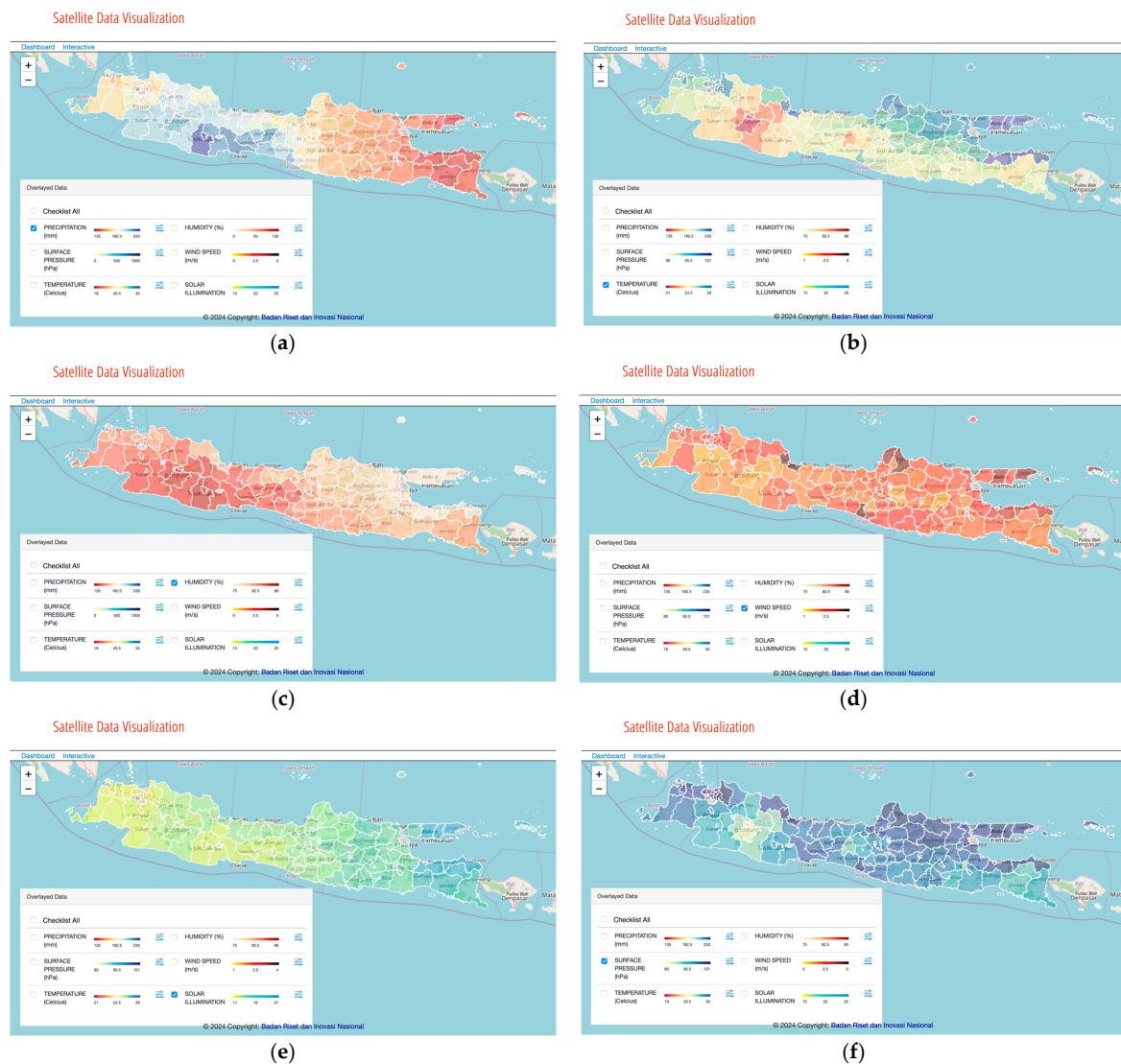


Figure 4. Spatial mapping of climate data in 119 districts/cities in Java Island, Indonesia: (a) rainfall, (b) air temperature, (c) humidity, (d) wind speed, (e) solar irradiation, (f) surface pressure.

3.5. Prediction Result of the ESDM

The estimation of prediction parameters in the ESDM was conducted using an R-Shiny web application. An estimated $\hat{\rho}$ value of 0.999 was obtained, producing an optimum spatial lag with a positive value ($\hat{\rho} > 0$) and indicating spatial lag dependence. It signified the influence of adjacent locations within the Java Island region on rainfall prediction data. It shows that the phenomenon of rainfall on Java Island has a positive spatial autocorrelation, meaning that if districts/cities on Java Island have high rainfall, then other districts/cities

around it have high rainfall as well or are not much different from the other districts/cities. The results of the parameter estimate calculation $\hat{\beta}_0$ and $\hat{\theta}$ are shown in Table 5.

Base on Table 5, the estimate $\hat{\beta}_0$ measures the direct impact of the exogenous variables on rainfall level in the same region, and the estimate $\hat{\theta}$ captures the spillover effects of the exogenous variables on rainfall level. Based on the estimate $\hat{\beta}_0$, we can obtain the estimate $\hat{\beta}$ from which different parameter estimates were obtained for each exogenous variable in 119 districts/cities. It explains that the ESDM produces different parameter estimates for each exogenous variable at each location due to the expansion of the exogenous variable matrix involving latitude and longitude information at each location. The highest effect on the rainfall is given by the surface pressure and the lowest is humidity. The ESDM equation for each location can be found in Appendix C. A visualization of rainfall prediction in 119 districts/cities of Java Island is shown in Figure 5.

Based on Figure 5, the highest monthly rainfall predictions, with values above 220 mm, were in West Java, such as in Ciamis, Tasikmalaya, and Pangandaran, while the lowest monthly rainfall predictions, with values below 140 mm, were in East Java, such as in Probolinggo, Probolinggo City, Situbondo, and Bondowoso. The proposed model resulted in a MAPE value of 1.956%, indicating very accurate prediction.

Table 5. Parameter-estimated value of SDM.

Coefficient	Parameter-Estimated Value	
	$\hat{\beta}_0$	$\hat{\theta}$
X_1 (air temperature)	$\hat{\beta}_{latitude}$	1.602
	$\hat{\beta}_{longitude}$	2.963
X_2 (humidity)	$\hat{\beta}_{latitude}$	0.197
	$\hat{\beta}_{longitude}$	0.427
X_3 (wind speed)	$\hat{\beta}_{latitude}$	0.838
	$\hat{\beta}_{longitude}$	0.143
X_4 (solar irradiation)	$\hat{\beta}_{latitude}$	6.598
	$\hat{\beta}_{longitude}$	-0.416
X_5 (surface pressure)	$\hat{\beta}_{latitude}$	-5.551
	$\hat{\beta}_{longitude}$	10.844

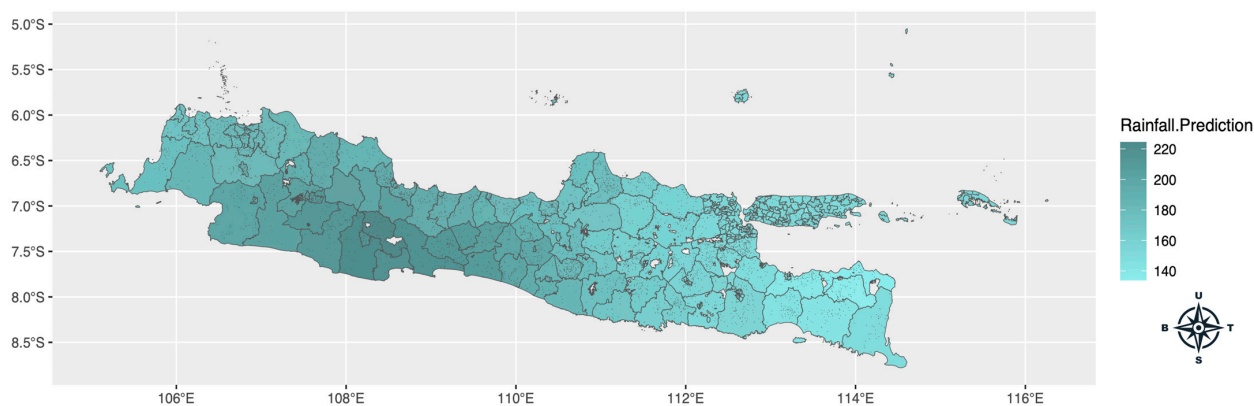


Figure 5. Spatial mapping of rainfall prediction in 119 districts/cities of Java Island.

4. Discussion

Data analytics life cycle methodology consists of six stages and is used to analyze big climate data sourced from NASA POWER. The initial step taken is formulating research

problems regarding the impact of rainfall on natural disasters. Problem identification includes spatial dependencies on climate variables that affect rainfall on the island of Java, Indonesia and initial hypotheses based on the theory that supports ESDMOK. Data preparation begins with collecting climate variable data containing location coordinate information, determining the observation time interval in the form of daily data, and selecting six climate variables: rainfall, air temperature, humidity, wind speed, solar irradiation, and surface pressure. Furthermore, the data pre-processing stage includes cleaning missing values, aggregating daily data to monthly data, and selecting duplicate data. Model planning integrates location data with climate variables as inputs in ESDMOK. Model development is one of the main objectives of our proposed technique for rainfall prediction. In the model-building stage, climate variables were predicted at 55 unobserved locations using the OK method based on 64 observed locations. The OK method uses the experimental semivariogram to fit the theoretical semivariogram model. The best theoretical semivariogram models for fitting are Gaussian and exponential models, with minimum SSE. Furthermore, based on the spatial autocorrelation test, spatial dependency exists in both rainfall and exogenous variables, contributing significantly to the accuracy of prediction results using ESDM. Communication of results gives very accurate prediction results, as shown by MAPE of 1.956%. The surface pressure variable has the largest influence on rainfall, and the smallest is wind speed. Furthermore, the post-processing stage is a visualization of spatial mapping for rainfall prediction and the interpretation of results to gain knowledge that can be used as a recommendation.

The ESDMOK was applied to predict rainfall in 119 districts/cities in Java Island, Indonesia, which is influenced by exogenous variables in the form of climate variables. Rainfall prediction is important in climate change in accordance with goal 13 of the Sustainable Development Goals (SDGs) concerns climate action. The summary of this research underscored that the level of rainfall in each region, based on the data from 119 districts/cities in Java Island, Indonesia, was significantly influenced by other climate-variable factors, such as air temperature, humidity, solar irradiation, wind speed and surface pressure [7]. The data complexity implies the need for a more effective technique. The incorporation of deep learning approaches and leveraging of big data should be considered to further enhance the prediction and analysis of rainfall in the study region.

5. Conclusions

In conclusion, this study proposes an ESDM and an interpolation technique with the OK method to calculate predictions at unobserved locations. ESDMOK can be used for rainfall prediction in spatial dependence on exogenous variables. Our proposed technique can identify spatial rainfall patterns, capture spatial dependence between observation units within the region and incorporate relevant exogenous variables; the model improves rainfall prediction accuracy. The surface pressure variable effects the most significant influence of exogenous variables on rainfall, and the smallest is wind speed.

The results of this model support disaster mitigation, water resources management, and infrastructure development that is resilient to natural disasters. The prediction results of the ESDMOK in all districts and cities in Java Island can be used as a recommendation by the Meteorology Climatology and Geophysics Agency (BMKG), Indonesia, agribusiness companies, and the general public in improving agricultural planning and planting seasons and providing climate information for the general public, especially related to rainfall in areas that have Monsoonal patterns.

6. Patents

Granted Copy Right: Copy Right for Computer Program, number 000484474.

Entitled "Application of RShiny Program for Ordinary Point Kriging Method on Rainfall Data in West Java", Ministry of Law and Human Rights of the Republic of Indonesia (Falah, A. N., Ruchjana, B. N., Abdullah, A. S., Rejito, J.), 2023. <https://annisanurfalah.shinyapps.io/Ordinary-Point-Kriging/> (accessed on 7 May 2024).

Author Contributions: Conceptualization, A.N.F. and Y.A.; methodology, A.N.F., Y.A., B.N.R. and E.H.; software, A.N.F., E.M. and S.B.S.; validation, T.H., R. and H.S.; formal analysis, A.N.F., Y.A., B.N.R., E.H., T.H., E.M., R., H.S. and S.B.S.; investigation, E.H., T.H., R. and H.S.; resources, A.N.F., E.H., T.H. and E.M.; data curation, E.H., T.H., R. and H.S.; writing—original draft preparation, A.N.F., Y.A. and B.N.R.; writing—review and editing, A.N.F., Y.A., B.N.R., E.H., T.H. and E.M.; visualization, A.N.F., E.M. and S.B.S.; supervision, Y.A., B.N.R. and E.H.; project administration, A.N.F. and Y.A.; funding acquisition, Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully thank Universitas Padjadjaran for providing financial support by a Postdoctoral Research Grant scheme with the contract number 2413/UN6.3.1/PT.00/2024 and Fundamental Research Grant with the contract number 4039/UN6.3.1/PT.00/2024 from the Ministry of Research, Technology and Higher Education Indonesia (Kemendikbudristek).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: We also thank the National Research and Innovation Agency (BRIN), Academic Leadership Grant Unpad 2024 and RISE_SMA project of European Union 2019–2024 for their assistance in conducting this research. Thanks to Atje Setiawan Abdullah, for the discussion and to all reviewers for their valuable comment and suggestion for this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Observed Locations (64 districts/cities) in Java Island, Indonesia.

Locations	Districts/Cities	Latitude	Longitude	Locations	Districts/Cities	Latitude	Longitude
1	Serang City	−6.11493	106.152694	33	Pemalang	−6.89214	109.378232
2	Pandeglang	−6.3092	106.1047	34	Brebes	−6.86278	109.037757
3	Tangerang City	−6.17139	106.640556	35	Wonosobo	−7.36793	109.900846
4	Tangerang Selatan City	−6.28578	106.712261	36	Magelang City	−7.48054	110.217695
5	Kepulauan Seribu	−5.6629	106.5683	37	Banjarnegara	−7.43482	109.566706
6	Jakarta Pusat	−6.17	106.82	38	Cilacap	−7.69801	109.024769
7	Bekasi City	−6.24159	106.992416	39	Sleman	−7.68167	110.32333
8	Bogor City	−6.59763	106.799568	40	Yogyakarta City	−7.80046	110.39128
9	Indramayu	−6.32758	108.324936	41	Kulon Progo	−7.8596	110.1579
10	Karawang	−6.32273	107.337579	42	Situbondo	−7.71667	114.05
11	Kuningan	−7.01381	108.570064	43	Probolinggo	−7.7353	113.4717
12	Kota Cirebon	−6.73725	108.550659	44	Sumenep	−7.02	113.87
13	Majalengka	−6.83638	108.227373	45	Jember	−8.1689	113.7022
14	Sumedang	−6.83812	107.927532	46	Bondowoso	−7.9404	113.9834
15	Garut	−7.22791	107.908699	47	Banyuwangi	−8.21861	114.366944
16	Cianjur	−6.82076	107.14296	48	Pasuruan City	−7.63333	112.9
17	Sukabumi	−7.06667	106.7	49	Probolinggo City	−7.75	113.216667
18	Bandung	−7.02525	107.51976	50	Lumajang	−8.13	113.22
19	Bandung City	−6.91486	107.608238	51	Sampang	−7.05	113.25
20	Pangandaran	−7.61506	108.498827	52	Mojokerto City	−7.47222	112.433611
21	Tasikmalaya City	−7.31956	108.202972	53	Surabaya City	−7.2458	112.7378
22	Sragen	−7.42028	111.023247	54	Blitar	−8.13333	112.25
23	Karanganyar	−7.60692	110.984515	55	Malang City	−7.98	112.62
24	Pati	−6.74905	111.037719	56	Nganjuk	−7.6	111.9333
25	Wonogiri	−7.79826	110.940606	57	Tuban	−6.9	112.1
26	Blora	−6.96874	111.418254	58	Bojonegoro	−7.15	111.88
27	Kudus	−6.80739	110.840369	59	Tulungagung	−8.0667	111.9
28	Semarang City	−7.00223	110.434226	60	Blitar City	−8.09861	112.165278
29	Surakarta City	−7.58104	110.826678	61	Kediri City	−7.81661	112.011917
30	Jepara	−6.57941	110.678479	62	Madiun City	−7.63	111.5231
31	Batang	−6.90668	109.733927	63	Pacitan	−8.13333	111.16667
32	Kendal	−6.93268	110.203074	64	Ponorogo	−7.8686	111.4619

Table A2. Unobserved Locations (55 districts/cities) in Java Island, Indonesia.

Locations	Districts/Cities	Latitude	Longitude	Locations	Districts/Cities	Latitude	Longitude
1	Serang	-6.15	106	29	Pekalongan	-6.89032	109.677
2	Cilegon City	-6.0204	106.0541	30	Tegal	-6.87027	109.1602
3	Lebak	-6.65	106.2167	31	Pekalongan City	-6.88981	109.6738
4	Tangerang	-6.3	106.5	32	Tegal City	-6.86728	109.1379
5	Jakarta Barat	-6.16717	106.7657	33	Magelang	-7.47986	110.2176
6	Jakarta Selatan	-6.25	106.8	34	Purworejo	-7.71297	110.01
7	Jakarta Timur	-6.2248	106.9011	35	Temanggung	-7.31343	110.1693
8	Jakarta Utara	-6.15225	106.8755	36	Banyumas	-7.47536	109.1615
9	Bekasi	-6.24667	107.1083	37	Kebumen	-7.6708	109.6614
10	Bogor	-6.59504	106.8166	38	Purbalingga	-7.38559	109.3617
11	Kota Depok	-6.38559	106.8307	39	Bantul	-7.88461	110.3341
12	Purwakarta	-6.53868	107.4499	40	Gunung Kidul	-7.96668	110.6026
13	Subang	-6.57159	107.7587	41	Pamekasan	-7.0667	113.5
14	Cirebon	-6.8	108.5667	42	Pasuruan	-7.73333	112.8333
15	Sukabumi City	-6.9237	106.9287	43	Jombang	-7.47	112.23
16	Bandung Barat	-6.8333	107.4833	44	Mojokerto	-7.55	112.5
17	Cimahi City	-6.89954	107.5339	45	Sidoarjo	-7.45303	112.7173
18	Banjar City	-7.37459	108.5582	46	Bangkalan	-7.02919	112.7461
19	Ciamis	-7.32622	108.3293	47	Gresik	-7.1933	112.553
20	Tasikmalaya	-7.5	108.1333	48	Lamongan	-7.12	112.42
21	Rembang	-6.70915	111.3421	49	Malang	-7.96688	112.6331
22	Demak	-6.89228	110.637	50	Batu City	-7.86667	112.5167
23	Grobogan	-7.02424	110.9187	51	Madiun	-7.61667	111.65
24	Semarang	-7.2486	110.4689	52	Trenggalek	-8.05	111.72
25	Boyolali	-7.51847	110.5932	53	Kediri	-7.83333	112.1667
26	Klaten	-7.74432	110.6678	54	Magetan	-7.65	111.37
27	Sukoharjo	-7.68017	110.8326	55	Ngawi	-7.4019	111.445
28	Salatiga City	-7.33102	110.51				

Appendix B

Theoretical semivariogram plot for climate variables, including rainfall, air temperature, humidity, solar irradiation, wind speed, and surface pressure.

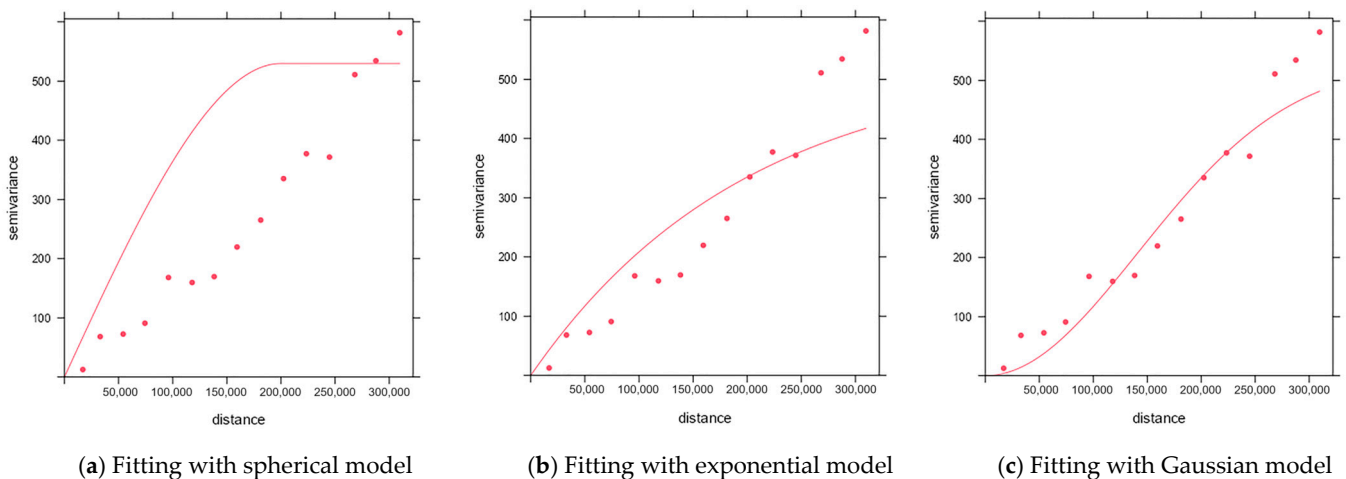
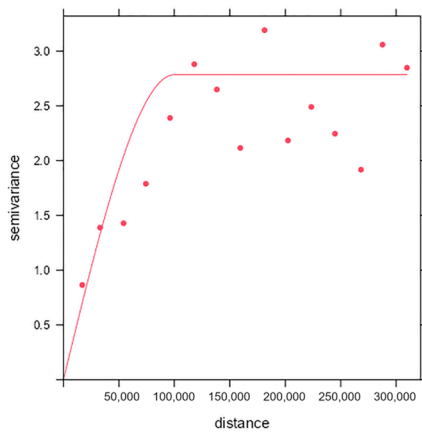
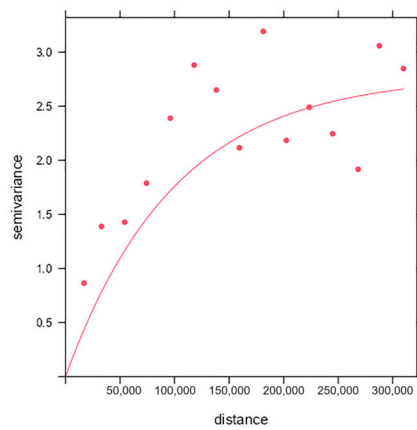


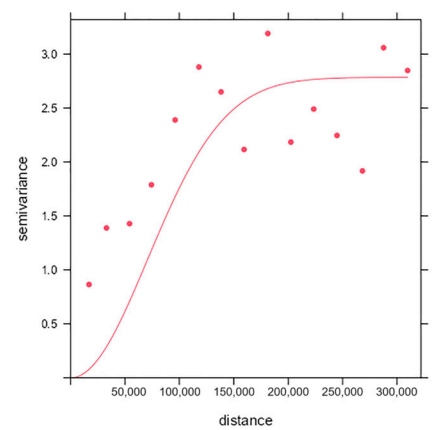
Figure A1. Theoretical semivariogram plot for rainfall.



(a) Fitting with spherical model

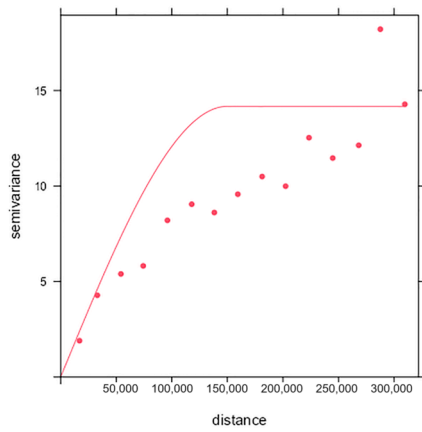


(b) Fitting with exponential model

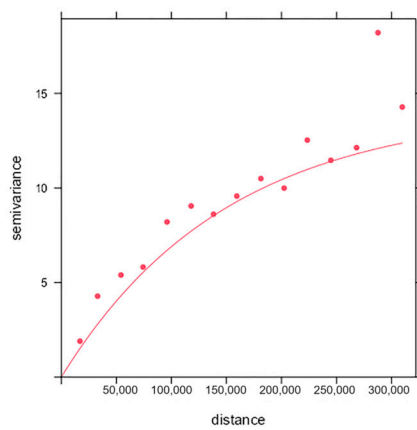


(c) Fitting with Gaussian model

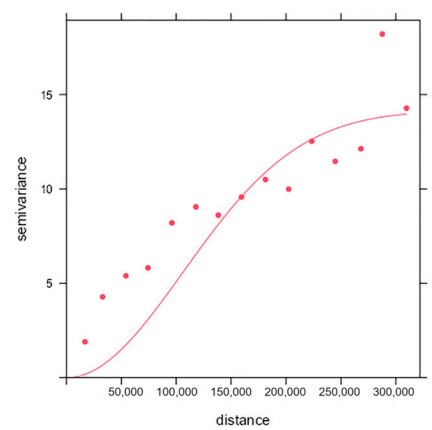
Figure A2. Theoretical semivariogram plot for air temperature.



(a) Fitting with spherical model

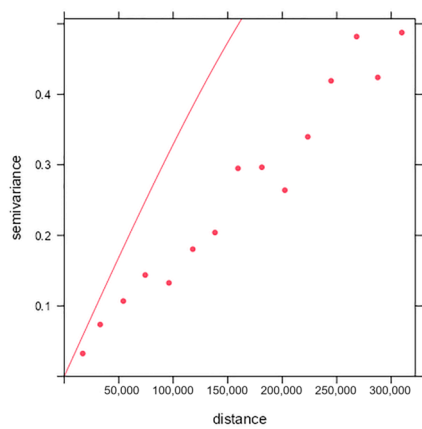


(b) Fitting with exponential model

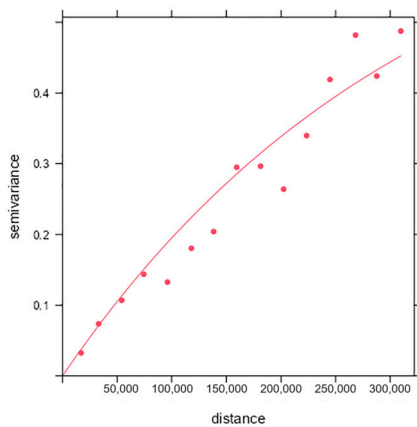


(c) Fitting with Gaussian model

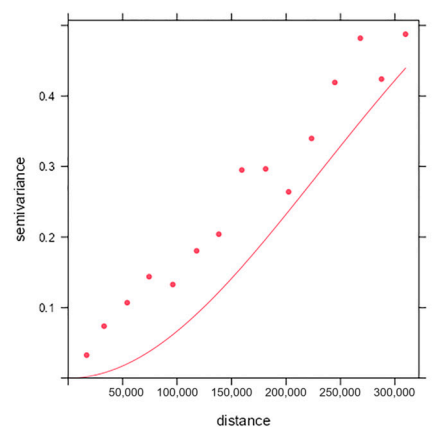
Figure A3. Theoretical semivariogram plot for humidity.



(a) Fitting with spherical model



(b) Fitting with exponential model



(c) Fitting with Gaussian model

Figure A4. Theoretical semivariogram plot for solar irradiation.

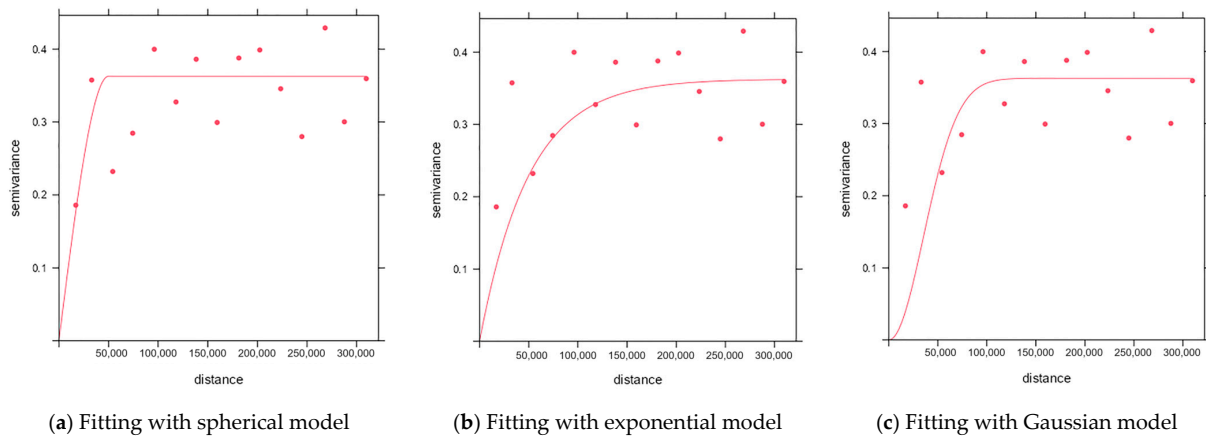


Figure A5. Theoretical semivariogram plot for wind speed.

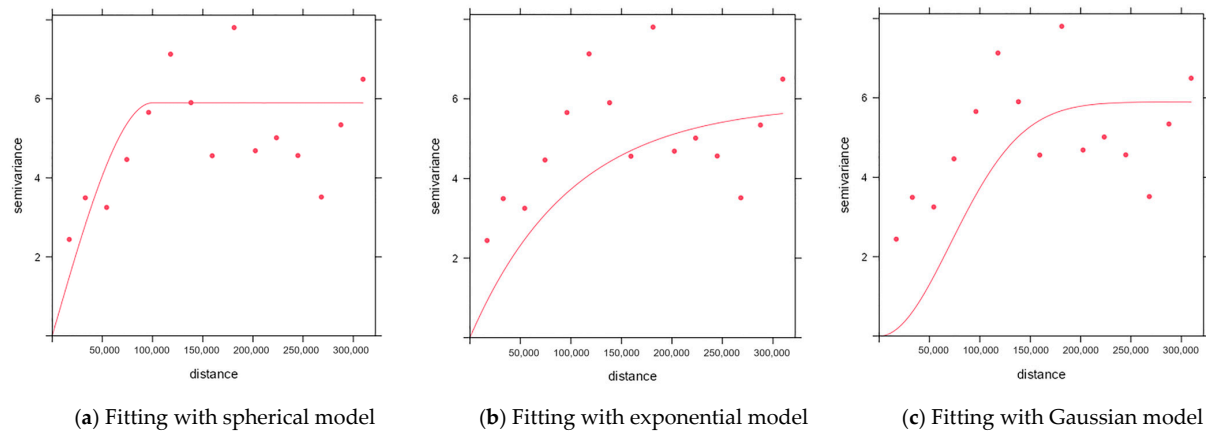


Figure A6. Theoretical semivariogram plot for surface pressure.

Appendix C

Table A3. The ESDM equations for predicting rainfall in 119 districts/cities of Java Island, Indonesia.

No	Locations	ESDM Equation for Predicting Rainfall
1	Serang City	$\hat{y}_{(s_1)} = 0.999 \sum_{i=1}^{64} w_{1i} y_{(s_1)} + 300.639 \times \mathbf{1}_{(s_1)} + 2.903X_1 - 0.340X_2 + 16.948X_3 - 5.394X_4 - 10.201X_5$ $- 9.262 \sum_{i=1}^{64} w_{1i} X_1 - 3.725 \sum_{i=1}^{64} w_{1i} X_2 - 9.497 \sum_{i=1}^{64} w_{1i} X_3 - 0.499 \sum_{i=1}^{64} w_{1i} X_4 + 3.015 \sum_{i=1}^{64} w_{1i} X_5$
2	Pandeglang	$\hat{y}_{(s_1)} = 0.999 \sum_{i=1}^{64} w_{1i} y_{(s_1)} + 300.639 \times \mathbf{1}_{(s_1)} - 2.135X_1 + 4.572X_2 + 4.572X_3 + 2.859X_4 + 22.605X_5$ $- 9.262 \sum_{i=1}^{64} w_{1i} X_1 - 3.725 \sum_{i=1}^{64} w_{1i} X_2 - 9.497 \sum_{i=1}^{64} w_{1i} X_3 - 0.499 \sum_{i=1}^{64} w_{1i} X_4 + 3.015 \sum_{i=1}^{64} w_{1i} X_5$
3	Tangerang City	$\hat{y}_{(s_1)} = 0.999 \sum_{i=1}^{64} w_{1i} y_{(s_1)} + 300.639 \times \mathbf{1}_{(s_1)} + 8.362X_1 - 4.063X_2 + 1.187X_3 + 5.062X_4 + 5.351X_5$ $- 9.262 \sum_{i=1}^{64} w_{1i} X_1 - 3.725 \sum_{i=1}^{64} w_{1i} X_2 - 9.497 \sum_{i=1}^{64} w_{1i} X_3 - 0.499 \sum_{i=1}^{64} w_{1i} X_4 + 3.015 \sum_{i=1}^{64} w_{1i} X_5$
...
119	Ponorogo	$\hat{y}_{(s_1)} = 0.999 \sum_{i=1}^{64} w_{1i} y_{(s_1)} + 300.639 \times \mathbf{1}_{(s_1)} + 3.881X_1 + 22.807X_2 - 1.230X_3 - 7.267X_4 + 1.912X_5$ $- 9.262 \sum_{i=1}^{64} w_{1i} X_1 - 3.725 \sum_{i=1}^{64} w_{1i} X_2 - 9.497 \sum_{i=1}^{64} w_{1i} X_3 - 0.499 \sum_{i=1}^{64} w_{1i} X_4 + 3.015 \sum_{i=1}^{64} w_{1i} X_5$

References

1. NASA Overview: Weather, Global Warming, and Climate Change. 2022. Available online: <https://science.nasa.gov/climate-change/what-is-climate-change/> (accessed on 28 February 2024).
2. Ditjenppi Dampak dan Fenomena Perubahan Iklim. 2022. Available online: <http://ditjenppi.menlhk.go.id/kcpi/index.php/info-iklim/dampak-fenomena-perubahan-iklim> (accessed on 7 March 2024).
3. BMKG Analisis Dinamika Atmosfer Dasarian III Mei 2022. Available online: <https://www.bmkg.go.id/iklim/dinamika-atmosfir.bmkg> (accessed on 5 April 2024).
4. SDGs Indonesia Sustainable Development Goals (SDGs)-Tujuan 13. 2021. Available online: <https://indonesia.un.org/id/sdgs/13/key-activities> (accessed on 5 April 2024).
5. Hatfield, G. Spatial statistics. In *Practical Mathematics for Precision Farming*; Wiley: Hoboken, NJ, USA, 2018; pp. 75–104. [CrossRef]
6. Stohlgren, T.J. Spatial Analysis and Modeling. In *Measuring Plant Diversity: Lessons from the Field*; Oxford University Press: Oxford, UK, 2007; pp. 254–270. [CrossRef]
7. Hermawan, E.; Lubis, S.W.; Harjana, T.; Purwaningsih, A.; Risyanto; Ridho, A.; Andarini, D.F.; Ratri, D.N.; Widyaningsih, R. Large-Scale Meteorological Drivers of the Extreme Precipitation Event and Devastating Floods of Early-February 2021 in Semarang, Central Java, Indonesia. *Atmosphere* **2022**, *13*, 1092. [CrossRef]
8. Falah, A.N.; Ruchjana, B.N.; Abdullah, A.S.; Rejito, J. The Hybrid Modeling of Spatial Autoregressive Exogenous Using Casetti's Model Approach for the Prediction of Rainfall. *Mathematics* **2023**, *11*, 3783. [CrossRef]
9. Andriyana, Y.; Falah, A.N.; Ruchjana, B.N.; Sulaiman, A.; Hermawan, E. Spatial Durbin Model with Expansion Using Casetti's Approach: A Case Study for Rainfall Prediction in Java Island, Indonesia. *Mathematics* **2024**, *12*, 2304. [CrossRef]
10. Abdullah, A.S.; Matoha, S.; Lubis, D.A.; Falah, A.N.; Jaya, I.G.N.M.; Hermawan, E.; Ruchjana, B.N. Implementation of Generalized Space Time Autoregressive (GSTAR)-Kriging model for predicting rainfall data at unobserved locations in West Java. *Appl. Math. Inf. Sci.* **2018**, *12*, 607–615. [CrossRef]
11. Falah, A.N.; Abdullah, A.S.; Parmikanti, K.; Ruchjana, B.N. Prediction of cadmium pollutant with ordinary point kriging method using Gstat-R. *AIP Conf. Proc.* **2017**, *1827*, 020019. [CrossRef]
12. Ruchjana, B.N.; Falah, A.N.; Abdullah, A.S. Application of the ordinary kriging method for prediction of the positive spread of COVID-19 in West Java. *J. Phys. Conf. Ser.* **2021**, *1722*, 012026. [CrossRef]
13. Gunawan, A.A.S.; Falah, A.N.; Faruk, A.; Lutero, D.S.; Ruchjana, B.N.; Abdullah, A.S. Spatial data mining for predicting of unobserved zinc pollutant using ordinary point Kriging. In Proceedings of the 2016 International Workshop on Big Data and Information Security (IWBIS), Jakarta, Indonesia, 18–19 October 2016; pp. 83–88. [CrossRef]
14. Gharaibeh, M.A.; Albalasmeh, A.A.; Moos, N.; Mohawesh, O.; Pratt, C.; El Hanandeh, A. A comparative analysis to forecast salinity and sodicity distributions using empirical Bayesian and disjunctive kriging in irrigated soils of the Jordan valley. *Environ. Earth Sci.* **2024**, *83*, 238. [CrossRef]
15. Falah, A.N.; Subartini, B.; Ruchjana, B.N. Application of universal kriging for prediction pollutant using GStat R. *J. Phys. Conf. Ser.* **2017**, *893*, 012022. [CrossRef]
16. Youkuo, C.; Yongguo, Y.; Wangwen, W. Coal seam thickness prediction based on least squares support vector machines and kriging method. *Electron. J. Geotech. Eng.* **2015**, *20*, 167–176.
17. Maria, E.; Budiman, E.; Taruk, M. Measure distance locating nearest public facilities using Haversine and Euclidean Methods. *J. Phys. Conf. Ser.* **2020**, *1450*, 012080. [CrossRef]
18. Montero, J.M.; Fernández-Avilés, G.; Mateu, J. *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*; Wiley: Hoboken, NJ, USA, 2012; ISBN 9781118762387.
19. Falah, A.N.; Hamid, N.; Rusyaman, E.; Abdullah, A.S.; Ruchjana, B.N. Implementation of Ordinary Co-Kriging method for prediction of coal quality variable at unobserved locations. *J. Phys. Conf. Ser.* **2021**, *1722*, 012076. [CrossRef]
20. Abdullah, A.S.; Hamid, N.; Falah, A.N.; Ruchjana, B.N. Prediction of spread shear strength of rock with ordinary point kriging method using GStat-R. *Appl. Math. Inf. Sci.* **2019**, *13*, 393–399. [CrossRef]
21. Lawrence, K.D.; Klimberg, R.K.; Lawrence, S.M. *Fundamentals of Forecasting Using Excel*; Industrial Press Inc.: New York, NY, USA, 2009; ISBN 083113335X.
22. Rahul, K.; Banyal, R.K. Data Life Cycle Management in Big Data Analytics. *Procedia Comput. Sci.* **2020**, *173*, 364–371. [CrossRef]
23. Munandar, D.; Ruchjana, B.N.; Abdullah, A.S.; Pardede, H.F. Literature Review on Integrating Generalized Space-Time Autoregressive Integrated Moving Average (GSTARIMA) and Deep Neural Networks in Machine Learning for Climate Forecasting. *Mathematics* **2023**, *11*, 2975. [CrossRef]
24. Stackhouse, P.J. NASA POWER Data Methodology. 2020. Available online: <https://power.larc.nasa.gov/> (accessed on 24 June 2024).
25. White, J.W.; Hoogenboom, G.; Stackhouse, P.W.; Hoell, J.M. Evaluation of NASA satellite- and assimilation model-derived long-term daily temperature data over the continental US. *Agric. For. Meteorol.* **2008**, *148*, 1574–1584. [CrossRef]
26. White, J.W.; Hoogenboom, G.; Wilkens, P.W.; Stackhouse, P.W.; Hoel, J.M. Evaluation of satellite-based, modeled-derived daily solar radiation data for the continental United States. *Agron. J.* **2011**, *103*, 1242–1251. [CrossRef]

27. Bai, J.; Chen, X.; Dobermann, A.; Yang, H.; Cassman, K.G.; Zhang, F. Evaluation of nasa satellite-and model-derived weather data for simulation of maize yield potential in China. *Agron. J.* **2010**, *102*, 9–16. [[CrossRef](#)]
28. Huffman, G.J.; Bolvin, D.T.; Braithwaite, D.; Hsu, K.; Joyce, R.; Xie, P. NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). Algorithm Theoretical Basis Document (ATBD) Version 4.4. *Natl. Aeronaut. Sp. Adm.* **2014**, 1–31. Available online: https://gpm.nasa.gov/sites/default/files/document_files/IMERG_ATBD_V5.2_0.pdf⁰https://pmm.nasa.gov/sites/default/files/document_files/IMERG_ATBD_%0AV4.4.pdf (accessed on 27 June 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.