Advances in Computer Science Research, volume 96
Proceedings of the International Conference on Mathematics, Geometry, Statistics, and Computation (IC-MaGeStiC 2021)

Statistical Downscaling Technique Using Response Based Unit Segmentation-Partial Least Square (REBUS-PLS) for Monthly Rainfall Forecasting

Izdihar Salsabila[1], Alfian Futuhul Hadi[1,*], I Made Tirta[1], Yuliani Setia Dewi[1], Firdaus Ubaidillah[2], Dian Anggraeni[1]

[1]*Data Science Research Group, Department of Mathematics, University of Jember, Indonesia*
[2]*Department of Mathematics, University of Jember, Jember 68121, Indonesia*
[*]*Corresponding author. Email: afhadi@unej.ac.id*

ABSTRACT
One of the newest forecasting techniques today is the Statistical Downscaling (SDs) technique. The SDs technique is a procedure for inferring high-resolution information from low-resolution variables. Forecasting rainfall using the SDs technique is to build a function that can predict the value of a response variable using predictor variables, for example, the variables in the Global Circular Model (GCM). In this study, forecasting will be carried out using the Partial Least Square (PLS) model and compared with the PLS model that has been time segmented namely the REBUS-PLS model. We use four latent variables consisting of three exogenous latent variables and one endogenous latent variable. The exogenous variable ξ_1 is precipitation, ξ_2 is air pressure, and ξ_3 is temperature, while the endogenous variable is monthly rainfall. The measurement model is a functional rule that describes the mathematical relationship between exogenous latent variables ξ_1, ξ_2, and ξ_3 with their corresponding manifests. After obtaining the structural model and measurement model, then parameter estimation is carried out. The PLS model obtained was then tested for the goodness of the model with several indicators, namely R^2, mean redundancy, and Goodness of Fit. The values obtained are 70.05%, 49.098%, and 76.11%. There are 4 segmentations which are segment 1 (33 months), segment 2 (29 months), segment 3 (50 months), and segment 4 (32 months). The validity and reliability tests were carried out again in each segment. Furthermore, the goodness of the model is also tested on each local model. The R-square values generated in segment 1, segment 2, segment 3, and segment 4 are 97.13%, 97.52%, 85.05%, and 91.38%. Overall, the PLS model has a smaller RMSE than the REBUS-PLS model at 25 observation stations. Meanwhile, at the other 52 observation stations, the accuracy of the REBUS-PLS model is better than the PLS model.

Keywords: *General Circulation Model (GCM), Statistical Downscaling (SDs), Partial Least Square (PLS), Response Based Unit Segmentation-Partial Least Square (REBUS-PLS).*

1. INTRODUCTION

Rainfall is one of the climate components that is often used as a reference, especially in agriculture. The erratic condition of rainfall fluctuation in recent years has caused agricultural planning to be suboptimal. Rainfall is a meteorological element with high variability in space and time scales, making it the most difficult to predict. Rainfall has the potential for both profitable and detrimental agriculture [1]. One area that has the potential to continue to develop its agricultural sector is Jember Regency. So far, the agricultural sector is a sector that has a reasonably significant role (leading sector) for the economy of Jember Regency. The latest data released by the Jember Regency Government show that around 41.73% of the total added value created in the economy of Jember Regency comes from the agricultural sector [2]. For that, we need a support system for agricultural activities in Jember Regency, one of which is the availability of information on current and future rainfall. This relates to the fact that cropping pattern planning will need to pay attention to the amount of rainfall in the future.

One of the most recent forecasting techniques is the statistical downscaling (SDs) technique. The SDs technique is a procedure for inferring high-resolution information from low-resolution variables. Forecasting rainfall using the SDs technique is to build a function that can predict the value of a response variable, namely rainfall using predictor variables, namely the variables in the global circular model (GCM). The Partial Least Square (PLS) model has been widely used in forecasting rainfall using statistical downscaling techniques. So far, the PLS model has been widely used in forecasting rainfall using statistical downscaling techniques. One of them is a study conducted by Estiningtyas and Wigena in 2011 [3]. The study compared the Principal Component Regression (PCR) and PLS models for forecasting rainfall under El Nino, La Nina, and normal conditions. Kurniawan states that REBUS-PLS can be done after the PLS-PM analysis finds characteristics in the quality of structural models that are not sufficiently representative, such as R^2 and Goodness of Fit (GoF) which indicate heterogeneity is not observed in the data [4]. Pratiwi, et al stated that the REBUS-PLS model was able to detect heterogeneity in the SEM-PLS model with the value of R^2 in each segment formed (local model) greater than the value of R^2 in the global model, which indicates that the local model is better than the global model [5].

In this research, forecasting will be carried out using the pls model and compared with the pls model which has been time segmented, hereinafter referred to as the rebus-pls model.

2. MATERIAL AND METHOD

2.1. Study Region

The research located was in Jember Regency. Jember Regency is one of the Regency in East Java that has an astronomical location with 113° 16′ 28" E to 114° 3′ 42" E longitude and 7° 59′6" S to 8° 33′56" S latitude. Jember Regency has an area of 3,293.34 km2 with a topographical character of fertile canyon plains in the middle and south and surrounded by mountains that extend the western and eastern borders [6]. Global Circular Model (GCM) data in this research was obtained from: http://climexp.knmi.nl/selectfield_cmip5.cgi. The area's boundaries used in this study are the latitude range of -21.25 to 3.75 and the longitude range of 101.25 to 126.25.

2.2. Data Description

In this study, two data were used, Global Circular Model (GCM) as explanatory variable and monthly rainfall data in Kabupaten Jember from January 2005 to December 2017 as the response. Monthly rainfall data in Jember Regency was obtained from 77 observation station points, each with coordinates longitude and latitude. The GCM data used are the precipitation variable (pr) in mm, air temperature (tas) in K units, and sea surface pressure (psl) in Pa units, from January 2005 to December 2017. In general, there are four latent variables. The variables used in this study are rainfall variables, precipitation variables, air temperature variables, and sea level pressure variables. The rainfall variable is composed of 77 manifest variables, each with historical rainfall data at each observation station. Meanwhile, the variables of precipitation, air temperature, and sea level pressure are each composed of 100 manifest variables, so that in total there are 377 manifest variables used in this study.

2.3. Method

2.3.1. Partial Least Square

SD modeling generally uses poorly conditioned covariates (large dimensions and has high correlation/multicollinearity). The model in this study is Partial Least Square (PLS) which can handle large-dimensional and multicollinearity problems. The first stage in the PLS model is to get a concept-based and structural model. The structural model is a design of the relationship between latent variables. We have used four latent variables consisting of three exogenous latent variables ξ and one endogenous latent variable η. The exogenous variable ξ_1 is precipitation, ξ_2 is air pressure, and ξ_3 is temperature, while the endogenous variable is monthly rainfall. The structural model of the four variables is first compiled in the form of a path matrix below:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

If the matrix A is described in the form of a path, then the path is obtained as shown in Figure 1 below:

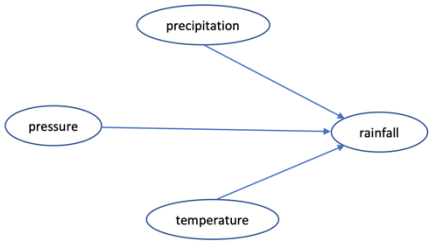

Figure 1 The Structural model trajectory.

Next, the structural model construction is carried out, so that it is obtained

$$\eta_i = [\xi_{1i} \quad \xi_{2i} \quad \xi_{3i}] \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} + b \qquad (1)$$

The next step is to design a measurement model. The design of the measurement model is the process of

determining the type of indicator of each latent variable. The equation below is a functional rule that describes the mathematical relationship between exogenous latent variables ξ_1 , ξ_2 , and ξ_3 with their corresponding manifestations, namely precipitation (pr), air pressure (psi), and temperature (tas). The measurement model is obtained as equation bellow:

$$\begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\end{bmatrix} = \begin{bmatrix}\lambda_{pr11} & \cdots & \lambda_{pr1010} & 0 & 0 & \cdots & \cdots & \cdots & 0\\ 0 & \cdots & 0 & \lambda_{psl11} & \cdots & \lambda_{psl1010} & 0 & \cdots & 0\\ 0 & 0 & \cdots & \cdots & \cdots & 0 & \lambda_{tas11} & \cdots & \lambda_{tas1010}\end{bmatrix}\begin{bmatrix}pr_{11}\\ \vdots\\ pr_{1010}\\ psl_{11}\\ \vdots\\ psl_{1010}\\ tas_{11}\\ \vdots\\ tas_{1010}\end{bmatrix} + \begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\end{bmatrix} \quad (2)$$

The next step is to determine the measurement model for each manifest Y. Manifest γ_1, γ_2, and so on are the rainfall variables at the first, second, and so on observation stations up to the 77th station. This equation will be the final stage of the calculation of the rainfall forecasting at 77 observation stations. Here is the measurement model of measurement for Y

$$\begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_{77}\end{bmatrix} = \eta\begin{bmatrix}\lambda_{Y1}\\ \lambda_{Y2}\\ \vdots\\ \lambda_{Y77}\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_{77}\end{bmatrix} \quad (3)$$

Next is the selection of the GCM variable output. One of the stages in PLS modeling is validity and reliability testing. A validity test is a test on each manifest whether it is feasible or not to be used as an explanation for the latent variables that it composes. The indicator of validity testing is to look at the loading on each manifest. The manifests included in the analysis are manifest with loading greater than 0.6. Meanwhile, the reliability test was carried out with the Cronbach Alpha indicator. The variable is explanatory that is consistent if the Cronbach Alpha value in each latent variable exceeds 0.5.

$$\begin{pmatrix}\xi_1\\ \xi_2\\ \xi_3\end{pmatrix} = \begin{bmatrix}\lambda_{pr12} & \cdots & \lambda_{pr109} & 0 & 0 & \cdots & \cdots & \cdots & 0\\ 0 & \cdots & 0 & \lambda_{psl11} & \cdots & \lambda_{psl109} & 0 & \cdots & 0\\ 0 & 0 & \cdots & \cdots & \cdots & 0 & \lambda_{tas11} & \cdots & \lambda_{tas108}\end{bmatrix}\begin{bmatrix}pr_{12}\\ \vdots\\ pr_{109}\\ psl_{11}\\ \vdots\\ psl_{109}\\ tas_{11}\\ \vdots\\ tas_{108}\end{bmatrix} + \begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\end{bmatrix} \quad (4)$$

If the initial data used consists of 377 variables with 100 precipitation variables, 100 air pressure variables, 100 temperature variables, and 77 rainfall variables, then after testing the validity and reliability tests, the data is reduced to 298 variables with 75 precipitation variables, 81 air pressure variables, 65 temperature variables, and 77 rainfall variables. So that the measurement model is rearranged into a new measurement model as in the equation (4).

The parameter estimation results in the PLS inner model are as follows

Table 1. Inner model parameter estimation

Coefficient	Estimation	Std.	p-value
Intercept	0.000	0.044	1.000
ξ_1	0.438	0.106	0.000
ξ_2	-0.903	0.102	0.000
ξ_3	0.352	0.090	0.000

Based on the parameter estimation results, the PLS structural model for forecasting monthly rainfall in Jember Regency is as follows

$$\begin{bmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_n\end{bmatrix} = \begin{bmatrix}1 & \xi_{11} & \xi_{21} & \xi_{31}\\ 1 & \xi_{12} & \xi_{22} & \xi_{32}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & \xi_{1n} & \xi_{2n} & \xi_{3n}\end{bmatrix}\begin{bmatrix}2{,}496\times10^{-15}\\ 0{,}4383\\ -0{,}9039\\ 0{,}3506\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix} \quad (5)$$

The structural model of rainfall forecasting using the PLS model is then tested using several test indicators to see how good the PLS model is in forecasting monthly rainfall. The test indicators used are the R-square value (R^2), the average redundancy, and the Goodness of Fit (GoF). The results of model testing on each indicator are as follows:

Table 2. Model goodness indicator

Variable	Type	R^2	*Mean*	GoF
$\boldsymbol{\xi_1}$	Exogenous	0.000	0.000	0.761
$\boldsymbol{\xi_2}$	Exogenous	0.000	0.000	
$\boldsymbol{\xi_3}$	Exogenous	0.000	0.000	
$\boldsymbol{\eta}$	Endogenous	0.701	0.491	

The model that is considered to have good quality is the model with GoF > 70%. The PLS model produced in this study has a GoF value of 76.11%.

2.3.2. Response Based Unit Segmentations (REBUS) on the PLS Model

The main purpose of implementing Response Based Unit Segmentations (REBUS) is to obtain several models in each of which the model can explain the diversity better than the PLS model. The initial stage in REBUS is choosing the number of segmentations. The number of segmentation in REBUS-PLS so far has been done using cluster analysis. The number of segmentation generated determines the number of local models formed in REBUS-PLS. The selection of the number of segmentation in this study uses the rainfall classification segmentation by Schmidt-Ferguson, which is as many as four segments. The results of segmentation on the resulting rainfall data are as shown in Table 3.

Table 3. Data segmentation results

No	Segment 1	Segment 2	Segment 3	Segment 4
1	Mar-05	Feb-05	Jan-05	Des-05
2	Jul-05	Apr-05	May-05	Mar-06
3	Aug-05	May-06	Jun-05	Des-07
4	Oct-05	Sep-06	Sep-05	Mar-08
5	Jan-06	Dec-06	Nov-05	Oct-08
6	Feb-06	Mar-07	Jun-06	Nov-08
7	Apr-06	Apr-07	Jul-06	Des-08
8	Feb-07	Jun-07	Aug-06	Apr-10
9	Feb-08	Nov-07	Oct-06	May-10
10	Jul-08	Jan-08	Nov-07	Jul-10

Table 4. Cronbach Alpha value in each segment

Segments	Latent Variable	Cronbach Alpha
Segment 1	Precipitation	0.997
	Air pressure	0.998
	Temperature	0.997
Segment 2	Precipitation	0.997
	Air pressure	0.998
	Temperature	0.997
Segment 3	Precipitation	0.997
	Air pressure	0.998
	Temperature	0.996
Segment 4	Precipitation	0.996
	Air pressure	0.998
	Temperature	0.994

Table 5. Parameter estimation value in each segment

Segment	Coefficient	Estimation	Std. Error	p-value
Segment 1	Intercept	0.000	0.030	1.000
	ξ_1	0.366	0.070	0.000
	ξ_2	-0.805	0.068	0.005
	ξ_3	0.550	0.062	0.000
Segment 2	Intercept	0.000	0.029	1.000
	ξ_1	0.321	0.070	0.000
	ξ_2	-0.654	0.057	0.000
	ξ_3	0.701	0.055	0.000
Segment 3	Intercept	0.000	0.051	1.000
	ξ_1	0.228	0.119	0.000
	ξ_2	-0.762	0.120	0.000
	ξ_3	0.397	0.112	0.000
Segment 4	Intercept	0.000	0.064	1.000
	ξ_1	0.349	0.132	0.015
	ξ_2	-0.109	0.159	0.000
	ξ_3	0.201	0.132	0.014

The Cronbach Alpha value based on Table 4 is quite large, exceeding 90%. This indicates that the segmentation division has been going well. Parameter estimation in the local model for each class can be seen in the following Table 5.

REBUS-PLS modeling requires retesting the validity and reliability of the data. It aims to see the consistency of latent and manifest variables in smaller data sizes. In the validity test, the loading value on the manifest of each segment is >0.6 or it can be said that the entire manifest value is eligible. The results of reliability test values are as shown in Table 4.

Based on the table, all exogenous latent variables have a significant effect on endogenous latent variables. Just as in the global model, the local model generated in each segment is tested for the goodness of the model. The results of model testing in each segment are as shown in Table 6.

Table 6. The goodness of the model on each segment

Segments	Variable	R^2	*Mean*	GoF
Segment 1	$\boldsymbol{\xi_1}$	0.000	0.000	0.878
	$\boldsymbol{\xi_2}$	0.000	0.000	
	$\boldsymbol{\xi_3}$	0.000	0.000	
	$\boldsymbol{\eta}$	0.972	0.616	
Segment 2	$\boldsymbol{\xi_1}$	0.000	0.000	0.864
	$\boldsymbol{\xi_2}$	0.000	0.000	
	$\boldsymbol{\xi_3}$	0.000	0.000	
	$\boldsymbol{\eta}$	0.975	0.556	
Segment 3	$\boldsymbol{\xi_1}$	0.000	0.000	0.828
	$\boldsymbol{\xi_2}$	0.000	0.000	
	$\boldsymbol{\xi_3}$	0.000	0.000	
	$\boldsymbol{\eta}$	0.850	0.703	
Segment 4	$\boldsymbol{\xi_1}$	0.000	0.000	0.845
	$\boldsymbol{\xi_2}$	0.000	0.000	
	$\boldsymbol{\xi_3}$	0.000	0.000	
	$\boldsymbol{\eta}$	0.912	0.620	

The average redundancy in each segment is 61.58%, 55.60%, 70.29%, and 61.97%. This indicates that the ability of exogenous variables to explain endogenous diversity in the PLS model in each segment is better than the global PLS model, which only has an average redundancy of 49.098%.

3. MODEL DEVELOPMENT

The PLS structural model for forecasting monthly rainfall in Jember Regency is as follows

$$\begin{bmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_n\end{bmatrix}=\begin{bmatrix}1 & \xi_{11} & \xi_{21} & \xi_{31}\\ 1 & \xi_{12} & \xi_{22} & \xi_{32}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & \xi_{1n} & \xi_{2n} & \xi_{3n}\end{bmatrix}\begin{bmatrix}2{,}496\times10^{-15}\\ 0{,}4383\\ -0{,}9039\\ 0{,}3506\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix} \quad (6)$$

The PLS model is generally divided into two parts, the inner model using the structural model and the outer model using the measurement model. The resulting PLS model can be used for forecast rainfall with GCM output data input of 221 variables according to the selected coordinates with 75 precipitation variables, 81 air pressure variables, and 65 temperature variables.

4. RESULT AND DISCUSSION

Previously, it was found that the local REBUS-PLS model in each segment could explain the diversity better than the global PLS model. Meanwhile, in forecasting rainfall, a model that is considered good is a model that can produce accurate forecasts. For this reason, forecasting is carried out on testing data as the final stage to test the ability of the PLS and REBUS-PLS models in forecasting monthly rainfall in the Jember Regency.

Table 7. Rainfall forecast results at 4 rainfall stations

Period	Puger			Cumedak		
	Actual	PLS	REBUS-PLS	Actual	PLS	REBUS-PLS
Jan17	262	227.48	202.41	546	519.24	474 91
Feb17	264	150.86	168.49	482	462.74	565.02
Mar17	184	127.23	146.49	492	481.22	578.14
Apr17	186	59.10	115.68	173	332.20	429.34
May17	72	64,81	68,73	94	181.77	211.91
Jun17	0	0.79	2,11	71	76.79	131,99
Jul17	7	52.04	8.45	0	55.19	26.37
Aug17	0	99.73	3.01	9	187.36	4.36
Sep17	2	15,27	16,53	97	148,22	63.04
Oct17	71	46.46	25.61	137	298.30	231.15
Nov17	255	62,44	53.65	335	414.89	415.79
Dec17	265	199.21	161.87	363	503.50	420.29
Error	-	85.41	77.73	-	102.61	101.18

Period	Lojejer			DAM.Klatakan		
	Actual	PLS	REBUS-PLS	Actual	PLS	REBUS-PLS
Jan17	221	269.37	238.24	490	454.52	439.29
Feb17	167	181,04	194.49	448	351.42	435,50
Mar17	138	139.54	170.69	413	346 81	464.19
Apr17	181	49.58	120.43	119	264.67	301.77
May17	41	84,09	84,92	108	87,65	130,22
Jun17	9	12,21	14.01	148	62.78	96.47
Jul17	2	78.24	12.76	46	112.11	122.22
Aug17	5	115,40	5,99	10	137,51	22.84
Sep17	3	8.32	12.79	48	98.31	47.03
Oct17	38	46.84	24.32	188	215.71	158.57
Nov17	236	67,23	64.84	417	349.06	327.21
Dec17	190	233.74	198.34	196	520.93	457.18
Error	-	76.47	55.98	-	78.01	82.06

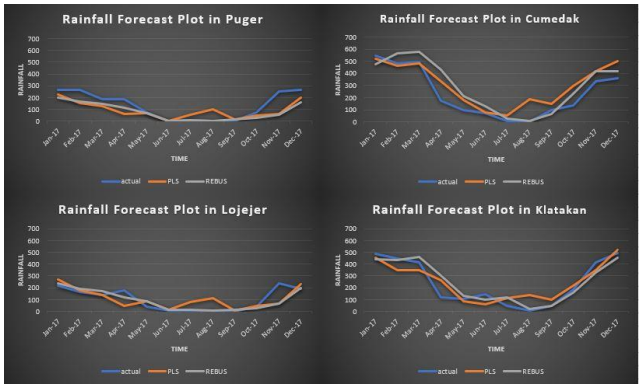

Figure 2 Rainfall forecast plot at 4 observation stations.

Rainfall plots at 4 observation stations are shown in Figure 2. If the RMSE values in both models at 77 observation stations are averaged, then the average RMSE value is 82.19 for the PLS model and 73.54 for the REBUS-PLS model. The PLS model has a smaller RMSE than the REBUS-PLS model at 25 observation stations. Meanwhile, at the other 52 observation stations, the accuracy of the REBUS-PLS model is better than that of the PLS model.

5. CONCLUSION

The rainfall forecasting model using the REBUS-PLS model can overcome the problem of diversity that is not well explained in the PLS model. Overall, the REBUS-PLS model has a better forecasting accuracy than the PLS model. The REBUS-PLS model has a smaller RMSE value than the PLS model at 52 observation stations. The PLS model produces a smaller RMSE compared to the REBUS-PLS model only at 25 observation stations.

REFERENCES

[1] B. Lakitan, Pengaruh Curah Hujan bagi Tanaman (in Indonesian), Universitas Indonesia, 2009.

[2] Bappeda Jawa Timur, Potensi Kabupaten dan Kota (in Indonesian), 2013, . http://bappeda.jatimprov.go.id/bappeda/wp-content/uploads/potensi-kab-kota-2013/kab-jember-2013.

[3] A. H. Wigena Regresi Kuadrat Terkecil Parsial Multi Respon untuk Statistical Downscaling (Multi Response Partial Least Square for Statistical Downscaling), Forum Statistika dan Komputasi, vol. 16, no. 2, pp. 12-15, Bogor, 2011.

[4] A. Kurniawan, Analisis Structural Equation Modeling Dengan Response-Based Units Segmentation Partial Least Square (Rebus-Pls) (in Indonesian), Thesis, Universitas Jember, 2018.

[5] F. S. Pratiwi, Sudarno, and S. Agus, Penerapan Response Based Unit Segmentation In Partial Least Square (Rebus-Pls) Untuk Analisis Dan Pengelompokan Wilayah (in Indonesian), Jurnal Gaussian, vol. 9, no. 3, pp. 364 – 375, 2020.

[6] R. Wilby, and T. Wigley, Downscaling General Circulation Model Output: A Review of Method ana Limitations, Progress in Physical Geography, vol. 4, pp. 530-548, 1997.