Determination of General Circulation Model Domain Using LASSO to Improve Rainfall Prediction Accuracy in West Java

Nanda Fadhli[1], Aji Hamim Wigena[2],Anik Djuraidah[3]
{nanda_senyo@apps.ipb.ac.id[1],aji_hw@apps.ipb.ac.id[2], anikdjuraidah@apps.ipb.ac.id[3]}

Department of Statistics, IPB University, Bogor, 16680, Indonesia, Phone/Fax (0251) 8624535[123]

Abstract. The Statistical downscaling technique has often been used to predict rainfall. This technique needsa domain of general circulation model (GCM) data. The selection of GCM domain is an important factor to improvepredictionaccuracy.The goal of this study is to determine the optimum domain. This study uses GCM data from CFSRv2 with gridresolution 2.5°×2.5°and local rainfall data in West Java. The GCM domain is determined basedon minimum correlation value of 0.3 between GCM data and local rainfall data. Correlations are calculated for the grid in the four directions of the compass with one grid as the reference that straightly above the local rainfall station. The domains are evaluated using the regression model with L1 (LASSO) regularization. The result showed that the optimum domain was 8×5 grids.

Keywords: GCM domain, statistical downscaling, LASSO regression

1 Introduction

West Java is faced with the risk of flooding (BMKG 2016) and drought (BMKG 2019) which destroy a thousand hectares of agricultural land. These are because of extremely high rainfall or low rainfall that faced by various societies group. Therefore, rainfall prediction information is needed.

General Circulation Model (GCM) can be used to obtain information for rainfall prediction [10]. The model produces data in the form of climate parameter values (such as precipitation) on the GCM grid. However, GCM grid resolution includes a large resolution so that precipitation on the grid cannot be used directly to describe rainfall in a local scale area with a smaller grid resolution [8].

Statistical downscaling (SDS) techniques can be used to obtain local scale information using global scale data. This technique requires a contiguous group of grids on GCM, called the GCM domain. The selection of GCM domains is an important factor in the SDS technique and will determine the prediction results [9]. Wigena (2006) used an 8×8 domain in Indramayuwhich resulted RMSEP about 63-98 and correlation about 0.50-0.76. Santri (2016) used an 8×8 domain in Indramayuwhich resulted RMSEP about 67-148 and correlation about 0.84-0.94.

This study aims to determine the optimum GCM domain in West Java. The GCM domain determination uses the concept of a minimum correlation of 0.3 (Garcia 2010) between the GCM grid data (predictor) and local rainfall data (response) for the grid located in the four

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290466

compass directions. The relationship closeness between response and predictors is one of the requirements in the SDS technique [3]. This domain is expected to improve the accuracy of rainfall prediction.

2 Data and Methodology

2.1 Data

This study uses two types of data. The predictor variables are GCM monthly precipitation data from 1981 until 2009 (348 months). These GCM data were generated from the *Climate Forecast System Reanalysis* (CFSR) model and could be downloaded from the website https://rda.ucar.edu/. The resolution of GCM data is 2.5° × 2.5° above the territory of Indonesia which located at 12°N to -15°S latitude and 90°E to 150°E longitude. The response variable is monthly rainfall data (mm) from 27 local rainfall stations in West Java which located at -7.78°S to -6.28°S latitude and 108.40°E to 107.87°E longitude. The rainfall data are from BMKG.

2.2Methods

The analysis procedure used R 3.5.2 software with the following steps:

1. Preparing data
   a. Transforming CFSR data from **netCDF** format into R dataframe using **ncdf4** package.
   b. Merging predictor variables (CFSR data) with response variables (local rainfall data).
   c. Cleaning the missing value so the data had a different length of data at each station, but in general, these data length is 348 months.
   d. Splitting data into modeling data and validation data. Modeling data are the entire month's data except validation data. While the validation data are data in the last 12 months.
   e. Exploring data by looking at the rainfall patterns.
2. Determining the optimum GCM domain for West Java.
   a. Determining the reference GCM grid which is the closest coordinate grid to the local rainfall station.
   b. Calculating the correlation value between local rainfall and GCM precipitation grid at the four compass directions.
   c. Selecting grids withminimum correlation value of 0.3.
   d. Forming a square or rectangular domain that includes the grids at point (2.c). Illustration of determining the GCM domain is shown in Figure 1.

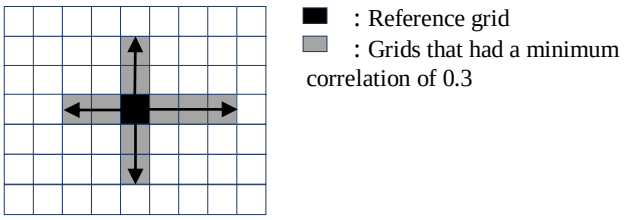

Fig. 1. Illustration of determining the GCM domain

e. Modeling the domain at point (2.d) using the LASSO regression with **glmnet** function on the R program, and looking for minimum lambda (λ) using the **cv.glmnet** function.
f. Determining the significance of grid using LASSO regression for each domain.
g. Merging all the significant grids from each domain into one large domain. If more than one grid intersects then only one grid includes in the domain.
h. Calculating the rainfall average for each type of land (such as low, medium, and plateau).
i. Calculating the correlation value between all grids in large domain and rainfall average for each type of land at point (h).
j. Selecting adjacent grids with correlation value ≥ 0.3, so that the optimum domain is formed.
k. Performing LASSO regression using the optimum domain for each land.
l. Calculating predictions and calculate RMSEP and correlation value for validation data.

2 Results and Discussion

Results. Data exploration is carried out on three local stations representing each type of land. The stations are Leles station (lowland 0-200 masl), Lengkong station (medium land 200-500 masl), and Sangiang station (plateau> 500 masl)[6]. Figure 2 shows that monthly rainfall increases with increasing altitude above sea level. Marpaung (2010) also revealed that land with an altitude of 600-1300 masl has the highest annual rainfall average compared to land with altitudes below 600 masl. U-shaped rainfall patterns are following the monsoon rainfall pattern in West Java.

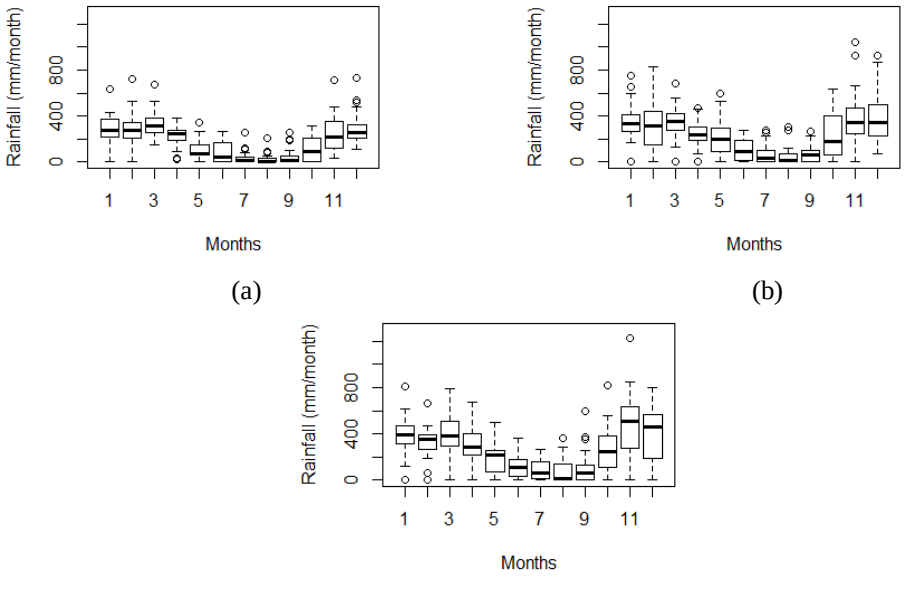

Fig. 2. Rainfall pattern of stasions: (a) Leles, (b) Lengkong, and (c) Sangiang

The results of each land (Table 1) shows that there are 17 stations with RMSEP between 37 and 100, while 10 stations with RMSEP more than 100. There are 19 stations having correlation between 0.8 to 0.96, and 8 stations having correlation more than 0.8. In general these domains results in relative small RMSEPs and high correlations.

Table 1. Domain evaluation results

Land	RMSEP		Correlation	
	37 to 100	100 to 156	0.71 to 0.8	0.8 to 0.96
Low	14	3	4	13
Medium	0	4	2	2
Plateau	3	3	2	4
All Stations	17	10	8	19

Large domains are shown in Figure 3. The domain still includes a number of insignificant grids, so the first and last rows in the domain are deleted, because more than 50% gridsare not significant. As the result, the size of large domain becomes 21×5.

	95	98	100	103	105	108	110	113	115	118	120	123	125	128	130	133	135	138	140	143	145
2.5	X78	X79	X80	X81	X82	X83	X84	X85	X86	X87	X88	X89	X90	X91	X92	X93	X94	X95	X96	X97	X98
0	X103	X104	X105	X106	X107	X108	X109	X110	X111	X112	X113	X114	X115	X116	X117	X118	X119	X120	X121	X122	X123
-2.5	X128	X129	X130	X131	X132	X133	X134	X135	X136	X137	X138	X139	X140	X141	X142	X143	X144	X145	X146	X147	X148
-5	X153	X154	X155	X156	X157	X158	X159	X160	X161	X162	X163	X164	X165	X166	X167	X168	X169	X170	X171	X172	X173
-7.5	X178	X179	X180	X181	X182	X183	X184	X185	X186	X187	X188	X189	X190	X191	X192	X193	X194	X195	X196	X197	X198
-10	X203	X204	X205	X206	X207	X208	X209	X210	X211	X212	X213	X214	X215	X216	X217	X218	X219	X220	X221	X222	X223
-13	X228	X229	X230	X231	X232	X233	X234	X235	X236	X237	X238	X239	X240	X241	X242	X243	X244	X245	X246	X247	X248
-15	X253	X254	X255	X256	X257	X258	X259	X260	X261	X262	X263	X264	X265	X266	X267	X268	X269	X270	X271	X272	X273

Fig. 3. Large Domain

Furthermore, the 21×5 domain is optimized to examine the relationship closeness between the predictor and response variables. This relationship is based on the correlation of the predictors in the domain grids and the response variables of each type of land. The optimum domain consists of grids with the correlation value> 0.3 and adjacent each other. Based on these criteria, the optimum domain size is 8×5.

The optimum domain is evaluated based on the values of RMSEP and correlation resulted from LASSO regression modeling. Table 2 shows that RMSEP in each land is relatively small and the correlation is relatively large. The result show that the domain size of 8×5 with grid size of 2.5°×2.5° is optimum. The optimum domain for West Java is shown in Figure 4.

Table 2.RMSEP and Correlation in Each Land

Land	Domain	RMSEP	Correlation
Low	8×5	23.61	0.99
Medium	8×5	75.78	0.96
Plateau	8×5	56.16	0.94

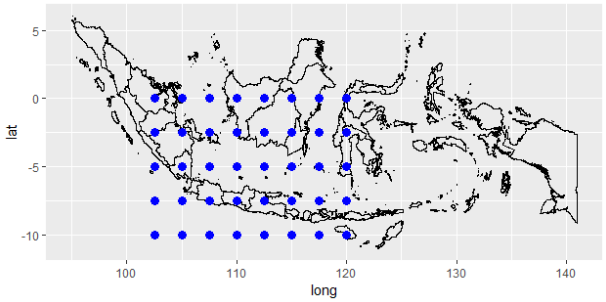

Fig. 4. The optimum GCM domain for West Java

3 Conclusion

Based on the results, it could be concluded that the optimum GCM domain for West Java is 8×5. This domain could improve the accuracy of monthly rainfall prediction compared to previous studies. The method of determining the GCM domain in this study also could be applied to other regions of Indonesia.

References

[1] BMKG [Badan Meteorologi, Klimatologi, dan Geofisika] : BMKG: Tren Hujan Terus Meningkat : https://jabarprov.go.id/index.php/news/20028/2016/11/16/BMKG-Trend-Curah-Hujan-Terus-Meningkat. (2019)

[2] BMKG [Badan Meteorologi, Klimatologi, dan Geofisika] : Potensi Kekeringan Meteorologis di Beberapa Wilayah di Indonesia. : https://www.bmkg.go.id/press-release/?p=potensi-kekeringan-meteorologis-di-beberapa-wilayah-di-indonesia&tag=&lang=ID (2019)

[3] Busuioc, A., Chen, D., Hellstro, C. : Performance of Statistical Downscaling Models in GCM Validation and Regional Climate Change Estimates : Application for Swedish Precipitation. International Journal of Climatology. 21 557–578 (2001)

[4] Garcia ,E. : A Tutorial on Correlation Coefficients. https://pdfs.semanticscholar.org/c3e1/095209d3f72ff66e07b8f3b152fab099edea.pdf. (2019)

[5] Marpaung, S. : PengaruhTopografiterhadap Curah HujanMusimandanTahunan di Provinsi Bali Berdasarkan Data ObservasiResolusi Tinggi. Prosiding Seminar PenerbangandanAntariksa. 104–110 (2010)

[6] Nuryanto, B., Priyatmojo, A., Hadisutrisno,B. :Pengaruh Tinggi TempatdanTipeTanamanPaditerhadapKeparahanPenyakitHawarPelepah :Jurnal Penelitian Pertanian Tanaman Pangan 33(1) 1. doi:10.21082/jpptp.v33n1.2014.p1-8 (2018)

[7] Santri, D. :Pemodelan Statistical Downscaling denganRegresiKuantilMenggunakan LASSO untukPendugaan Curah HujanEkstrim. Bogor (ID): Institut Pertanian Bogor(2016)

[8] Wigena, A.H.: Pemodelan Statistical Downscaling DenganRegresi Projection Pursuit UntukPeramalan Curah HujanBulanan (Kasus Curah hujanbulanan di Indramayu).Bogor (ID): Institut Pertanian Bogor (2006)

[9] Wilby, R.L., Wigley, T.M.L.: Precipitation predictors for downscaling: Observed and general circulation model relationships. International Journal of Climatology. 20(6) 641–661. doi:10.1002/(SICI)1097-0088(200005)20:6<641::AID-JOC501>3.0.CO;2-1 (2000)

[10]Zorita, E., Von Storch, H.: The Analog Method as a Simple Statistical Downscaling Technique : Comparison with More Complicated Methods. Journal of Climate. 2474–2489. doi:10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2 (1999)