Combining Result of PCR and PLSR Statistical Downscaling with Quadratic Optimization for Improving Estimation

Praditya Puspitaninggar [1], Aji Hamim Wigena[2], Anwar Fitrianto[3]
{puspitaninggar_p@apps.ipb.ac.id[1], ajiwigena@ymail.com[2], anwarstat@gmail.com}

Statistics, IPB University, Bogor, 16680, Indonesia[1], Statistics, IPB University, Bogor, 16680, Indonesia [2], Statistics, IPB University, Bogor, 16680, Indonesia [3]

Abstract. Statistical downscaling (SDS) using principal component regression (PCR) and partial least square regression (PLSR) used to estimate rainfall. The accuracy measurement of these models was done by calculating root mean square error of prediction (RMSEP) value. The smaller RMSEP value, the closer estimate to actual rainfall. The RMSEP value of PCR and PLSR models tend to be large, so it needs a method to improve the accuracy of estimation. Improving accuracy of rainfall estimate was done by combining the two estimates with the SDS models and minimizing the combined RMSEP as an objective function with or without constraints. The optimization with constraints resulted in the combined RMSEPs larger than the RMSEPs of the two models on few datasets, while the optimization without constraints resulted in the combined RMSEPs smaller than the RMSEPs of the two models on all datasets. The combined rainfall estimate was better than rainfall estimates forming it.

Keywords: Combined forecast, partial least square regression, principal component regression, statistical downscaling, quadratic optimization.

1 Introduction

Rainfall can involve many of natural disaster such as floods, landslides, and dryness. Necessary to avoid many victim and material losses that caused by the natural disaster is by develop a model to predict rainfall that have minimum error. A method that can used for estimating rainfall is Statistical Downscaling (SDS). This model is a regression between local rainfall and global circulation model (GCM) data. GCM data is large dimension data and there are multicollinearity problems, so modelling SDS can be done by principle component regression (PCR) and partial least square regression (PLSR) that have done by Sari (2015) and Matulessy (2015) and had the root mean square error prediction (RMSEPs) are 74.48 and 72.17 respectively. The smallest RMSEPs indicate a better model. Combining the result of modelling is a method to improving the accuracy. Osman et al. (2016) develops a model to improve the prediction accuracy by combined sources or methods. The simple combined modelling result can be done with combining SDSM method, multiple linear regression (MLR), and Gamma generalized linear model (GLM) based on mean of prediction. A research that concern with combining prediction was done by Wigena et al. (2014) that weighted summarizing. The optimum weight that used is iterative procedure and used linear regression without intercept. The alternative procedure for calculate the optimum weight can be done by

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290531

lagrange multipliers method. This method better because no need trial and error to calculate the optimum weight (Vertisa 2014). This research is combining the PCR and PLSR result by quadratic programming optimizing to improve prediction accuracy.

2 Objectives

In this research, we have to improve the accuracy of rainfall estimates with combine two methods of statistical downscaling. These are principal components regression and partial least square regression. This new method is combine the result of these two methods using quadratic programming optimization.

3 Materials

The materials for this research is output precipitation GCM climate model intercomparison project (CMIP5) data for X variable and rainfall data from BMKG Indramayu (ZOM 79000) 1981-2013 in stasiun Krengkel, Sukadana, Karangkendal, and Gegesik for Y variable. GCM domain is a square 8×8 grid (2.5° × 2.5° for each grid) in 98.75°BT - 116.25°BT dan 16.25°LS - 1.25°LU above the region of Indramayu. For training data are 1981-2012 and 2013 as testing data.

4 Methods

Analysis method that used for this research is combining prediction based on PCR and PLSR model by quadratic optimizing with and without constraint that involved RMSEP. Data analyze for this research is done by pls, quadprog, ggplot2, and caret package in R 3.5.2 software that follow:

1. Data exploratory to identification general description from rainfall in Indramayu 1981-2013,
2. Ordinary least square (OLS) regression for calculate the variance inflation factor (VIF) that are an instrument to identification the multicollinearity.
3. Predict the rainfall by PCR model that follow these algorithms:
    a. Defined X correlation matrix from training data
    b. Calculate the eigen vector from matrix correlation
    c. Form 64 principle components based on linear combination of eigen vector and matrix X
    d. Define how many principle components that used for PCR by cross-validation to get the minimum RMSE
    e. Validate the model with testing data
4. Predict the rainfall by PLSR
5. Define the constrained optimum weight based the result of RMSEP value from PCR and PLSR models that follow these algorithms:
    a. Define the objective function by:

$$f(w) = \min \text{RMSEP}_{\text{w}} = \sqrt{\frac{\sum_{i=1}^{t}\left(y_i - \hat{y}_{gab_i}\right)^2}{t}}$$

$$= \min\Big(\sum_{i=1}^{t} y_i^2 - 2w_1 \sum_{i=1}^{t}(y_i^2.\hat{y}_{RKU_i}) - 2w_2 \sum_{i=1}^{t}(y_i^2.\hat{y}_{RKTP_i}) + w_1^2 \sum_{t}^{t} \hat{y}_{RKU_i}^2 + 2w_1 w_2 \sum_{i=1}^{t}(\hat{y}_{RKU_i}^2.\hat{y}_{RKTP_i}^2) + w_2^2 \sum_{i=1}^{t} \hat{y}_{RKTP_i}^2$$

$$f(w) = aw_1^2 + bw_2^2 + 2cw_1w_2 - 2dw_1 - 2ew_2 + f$$

Information:

$a = \sum \hat{y}_{RKU}^2$

$b = \sum \hat{y}_{RKTP}^2$

$c = \sum \hat{y}_{RKU}.\hat{y}_{RKTP}$

$d = \sum y_i.\hat{y}_{RKU}$

$e = \sum yi.\hat{y}_{RKTP}$

$f = \sum yi^2$

follow quadratic form:

$$f(w) = \frac{1}{2}\boldsymbol{w}^T 2\boldsymbol{Q}\boldsymbol{w} - \boldsymbol{b}^T\boldsymbol{w} + c$$

with:

$$\boldsymbol{A} = \begin{bmatrix} \sum \hat{y}_{RKU}^2 & \sum \hat{y}_{RKU}.\hat{y}_{RKTP} \\ \sum \hat{y}_{RKU}.\hat{y}_{RKTP} & \sum \hat{y}_{RKTP}^2 \end{bmatrix}$$

$$\boldsymbol{b} = [\sum y_i.\hat{y}_{RKU} \quad \sum yi.\hat{y}_{RKTP}]$$

b. Constraint function $w_1 + w_2 = 1$ with matrix form $\boldsymbol{C}\boldsymbol{w} = \boldsymbol{d}, \boldsymbol{C} = [1 \quad 1]$ and $\boldsymbol{d} = [1]$

c. Lagrange function $L(w_i, \lambda) = \frac{1}{2}\boldsymbol{w}^T 2\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b} + c + \lambda[\boldsymbol{C}\boldsymbol{w} - \boldsymbol{d}]$

d. Optimize with $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}^T \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix}\begin{bmatrix} \boldsymbol{w} \\ \lambda \end{bmatrix} = \begin{bmatrix} -(-\boldsymbol{b}) \\ d \end{bmatrix}$ and gets $\begin{bmatrix} \boldsymbol{w} \\ \lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}^T \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix}^{-1}\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{d} \end{bmatrix}$

6. Calculate the constrained combine prediction ($\hat{y}_{com}$) with optimum weight (w_1) and (w_2) from PCR prediction ($\hat{y}_{PCR}$) and PLSR prediction ($\hat{y}_{PLSR}$) that follow equation:

$$\hat{y}_{com} = w_1.\hat{y}_{PCR} + w_2.\hat{y}_{PLSR}$$

7. Calculate the $\text{RMSEP}_{\text{com.con}} = \sqrt{\frac{(y_i - \hat{y}_{com})^2}{t}}$
8. Unconstrained optimization from function $f(w_1, w_2) = aw_1^2 + bw_2^2 + 2cw_1w_2 - 2dw_1 - 2ew_2 + f$
9. Calculate the unconstrained combine prediction ($\hat{y}_{com}$) with optimum weight (w_1) and (w_2) from PCR prediction ($\hat{y}_{PCR}$) and PLSR prediction ($\hat{y}_{PLSR}$) that follow equation:

$$\hat{y}_{com} = w_1.\hat{y}_{PCR} + w_2.\hat{y}_{PLSR}$$

10. Calculate the $\text{RMSEP}_{\text{com.uncon}} = \sqrt{\frac{(y_i - \hat{y}_{com})^2}{t}}$
11. Compare the result of constrained and unconstrained estimates

5 Result

5.1 Data Description

During 33 years, rainfall in Indramayu is illustrated by a boxplot in **Figure 1**. In general, the rainfall in Indramayu can be seen as having a U-pattern. January, February and December are the months with the highest average rainfall values that up to more than 200 mm/month from 1981-2013 so that it is called wet months. While the lowest rainfall occurs in July, August and September or is that is called dry months. The remaining six months are months of seasonal change or transition. **Figure 1** below also shows that in certain months there is extreme rainfall described by outliers in each month such as March to June, September, November and December.

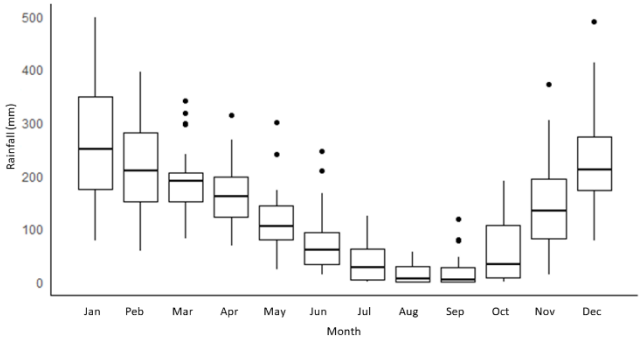

Fig. 1. Boxplot of Indramayu monthly rainfall 1981-2013

5.2 Principle Component Regression

PCR analysis starts with define how many principles component (PC) that used with principle component analysis based on cross-validation to get the minimum RMSE. **Figure 2** shown that best model is model with four PC. Model with one PC has RMSE greater than 90, model with two PC has RMSE greater than 70, and model with more than five PC deliver RMSE greater than 70.

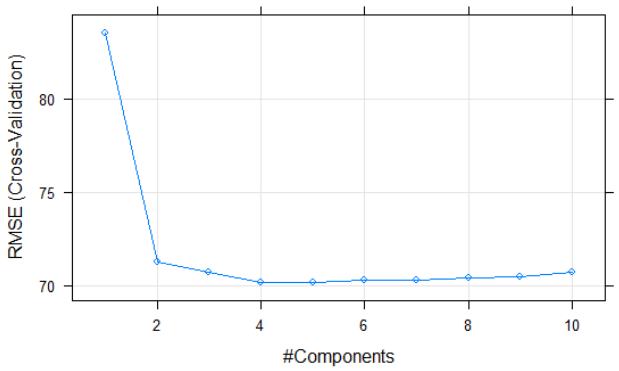

Fig. 2. Cross-validation of RMSE PCR model

The fourth PC are have biggest variance that is 0.693. Variance proportion that can be explained by the fourth PC is more than 80%. Scree plot in **Figure 3** shown that the variance of PC1 and PC2 are rapidly quiet, but PC1 only explained 69.3% variance of X. Furthermore, in this research is used PC1, PC2, PC3, and PC4 that have eigen value more than one and can explained variance of X until 96.6%. This is shown by **Table 1**.

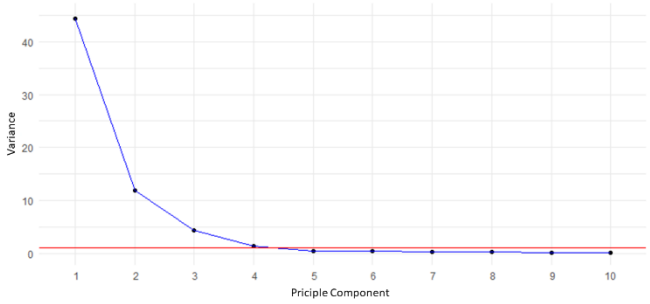

Fig. 3. Scree plot of principle component variance

Table 1. Variance proportion of principle components.

	PC1	PC2	PC3	PC4	PC5	PC6	…	PC64
Standard deviation	6.659	3.436	2.082	1.167	0.673	0.616	…	0.013
Variance propotion	0.693	0.184	0.068	0.021	0.007	0.006	…	0.000
Proportion cumulative	0.693	0.877	0.945	0.966	0.973	0.979	…	1.000

Validation was done by estimate Indramayu rainfall for January until December 2013 with PCR model. The estimates for this model and the actual value of Indramayu rainfall is shown by **Figure 4.** This figure shown that generally rainfall estimates based on PCR with four PC has same pattern with the actual value. The rainfall is high in wet months and low in dry months. For several months such as February, the estimates is different with the actual rainfall. Rainfall estimates with PCR model has RMSEP equal to 72.74.

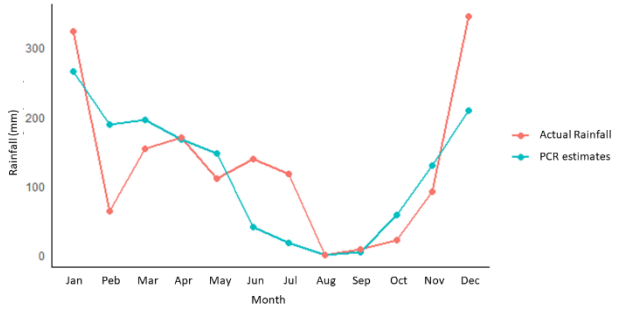

Fig. 4. Rainfall estimates for 2013 with PCR model

5.3 Partial Least Square Regression

The best PLSR model is model with four components. This is proven by cross-validation of RMSE in **Figure 5**. PLSR model that used four components can explained 99.37% variance of Y and 94.75% variance of X. This is shown by **Table 2**. This model is a good model because can explained most of the variance of X.

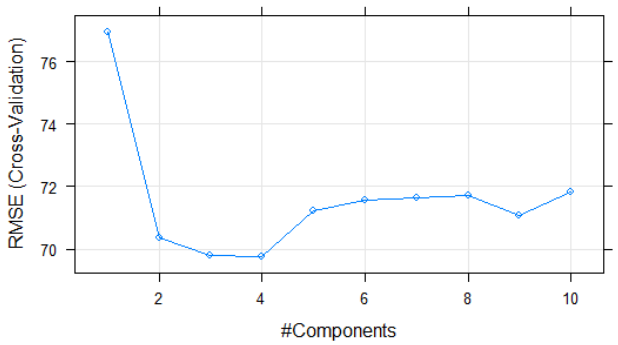

Fig. 5. Cross-validation of RMSE PLSR model

Table 2. Percent of variance of X and Y that explained PLSR components

Variable	1 Comp	2 Comp	3 Comp	4 Comp	5 Comp
X	67.25	87.02	90.58	94.75	96.69
Y	49.87	62.78	84.90	99.37	99.89

Rainfall estimates for Indramayu with PLSR model was done for January-December 2013. The estimates with PLSR model give RMSEP that greater than RMSEP of PCR model that is equal to 74.54. Generally, rainfall estimates for PLSR model that shown by **Figure 6** can follow the actual rainfall and seems similar with the estimates of PCR model. Deviation between estimates and actual rainfall in February seems big.

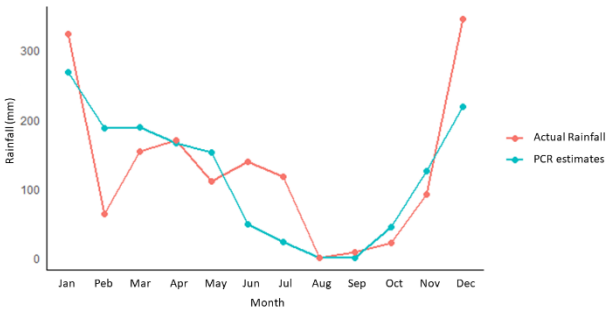

Fig. 6. Estimates rainfall for 2013 with PLSR model

5.4 Combining the Result of PCR and PLSR Model

Rainfall estimates was done by combining the result of PCR and PLSR model with constrained and unconstrained optimization. The constrained optimization model was analyzed with quadratic optimization for minimize the

RMSEP. This procedure obtained weight that is equal to 0.99 for PCR estimates and 0.01 for PLSR estimates. These weights have difference for these two models, so the combined estimation give a pattern that similar to pattern of PCR and PLSR estimates. This combined model has RMSEP equal to 74.51 that is less than the RMSEP of PLSR model but greater than the RMSEP of PCR model. The estimates for this combined model is shown by **Figure 7**.

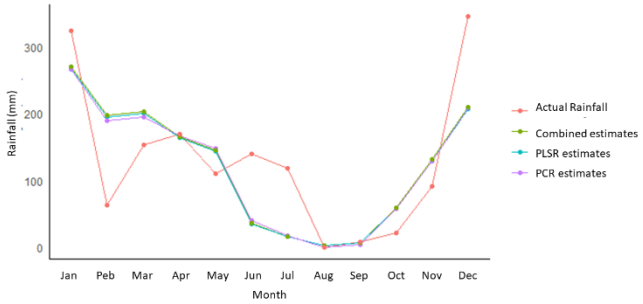

Fig. 7. Constrained combined estimates for 2013

$$a + b = c \,. \tag{1}$$

Goodness of a model can be shown by several combination of difference testing data and training data. Combined model from PCR and PLSR model using constrained quadratic optimization has the RMSEP that similar with both of developer model. Tabel 3 shown that combined RMSEP from several datasets.

Table 3. Weight and RMSEP of combined model with constrained optimization

Periods	Weight of RKU	Weight of RKTP	RMSEP of RKU	RMSEP of RKTP	RMSEP of Combined	Correlation Coefficient
1981 – 2012	**0.99**	**0.01**	**72.75**	**74.54**	**74.51**	**0.73**
1981 – 2011	0.27	0.73	47.01	47.92	24.61	0.97
1981 – 2010	0.17	0.83	64.11	65.34	59.82	0.72
1981 – 2009	0.93	0.07	76.12	77.73	76.81	0.64
1981 – 2008	0.17	0.83	61.39	62.64	56.95	0.77
1981 – 2007	0.76	0.24	75.66	73.72	63.74	0.93
1981 – 2006	0.99	0.01	52.18	53.38	53.34	0.86
1981 – 2005	0.87	0.13	80.86	80.55	78.18	0.86
1981 – 2004	0.08	0.92	52.99	54.11	52.68	0.81
1981 – 2003	0.09	0.91	76.40	77.50	76.30	0.75

Unconstrained optimization was done by optimizing RMSEP without a limitation. Combined PCR and PLSR model with optimize unconstrained function have good estimates. Weight for PCR estimates is 13.83 and -12.79 for PLSR estimates. These weights is totally different with weights in constrained optimization. The estimates by using these weights

seems good because obtained RMSEP that is equal to 58.43. The estimates also have a pattern that can follow the actual rainfall of Indramayu. This model can estimate rainfall of February better. The pattern for this estimation is shown by **Figure 8.**

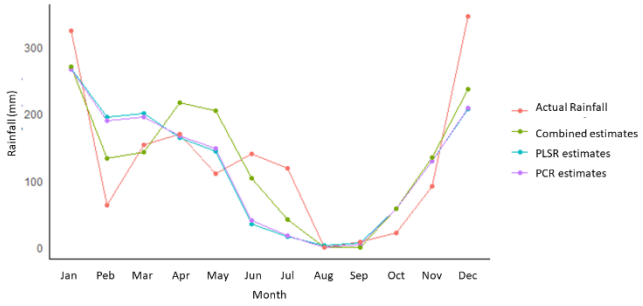

Fig. 8. Unconstrained combined estimates for 2013

Rainfall estimates for 2013 from several datasets calculated to prove the consistency of this model. Analysis for ten datasets was done and have smaller RMSEP than constrained optimization. **Table 4** shown the weights and RMSEPs for ten datasets. Overall, the RMSEPs is smaller than the developer model. The standard deviation of the RMSEP is equal to 14.89 that smaller than standard deviation of the RMSEP of constrained optimization that is 15.60. It is shown that this model is a consistent model and can be used for other datasets. The actual rainfall seems have a strong relationship with this combined estimation that has correlation coefficient is equal to 0.70.

Table 4. Weight and RMSEP of combined model with unconstrained optimization

Periods	Weight of RKU	Weight of RKTP	RMSEP of RKU	RMSEP of RKTP	RMSEP of Combined	Correlation Coefficient
1981 – 2012	**13.83**	**-12.79**	**72.75**	**74.54**	**58.43**	**0.76**
1981 – 2011	1.50	-0.76	47.01	47.92	23.57	0.81
1981 – 2010	3.18	-2.32	64.11	65.34	58.40	0.86
1981 – 2009	6.47	-5.36	76.12	77.73	70.47	0.89
1981 – 2008	3.92	-3.11	61.39	62.64	53.35	0.92
1981 – 2007	-5.48	6.72	75.66	73.72	57.81	0.80
1981 – 2006	8.58	-7.50	52.18	53.38	47.27	0.70
1981 – 2005	-0.29	1.42	80.86	80.55	78.16	0.73
1981 – 2004	1.96	-1.01	52.99	54.11	51.81	0.96
1981 – 2003	2.19	-1.25	76.40	77.50	75.24	0.83

6 Conclusion

The results of statistical downscaling modeling with principal component regression and partial least square regression give almost the same RMSEP value. Likewise, the estimated monthly rainfall values of the two models give a similar pattern.

The two patterns of the model are also able to follow the actual rainfall pattern, namely high during the wet month, low during the dry month, and the rest have an upward or downward transition pattern. However, the two patterns of the plot had considerable deviations in February.

Constrained integration can improve accuracy because it provides a smaller RMSEP than the two builder models. Estimation of rainfall with the combination of constraints is inconsistent and not good. The unconstrained combination produces a consistent and very good value of rainfall estimates. This alleged value has a strong attachment to the actual rainfall. The correlation coefficient between actual rainfall and the estimated value of the merger with the optimization of the non-constrained function is quite good, which means there is a close relationship between the two.

7 References

Matulessy ER. 2015. Regresi kuantil dengan kuadrat terkecil parsial dalam statistical downscaling untuk pendugaan curah hujan ekstrem [tesis]. Bogor (ID): Institut Pertanian Bogor.

Osman YZ, Mawada E, Abdellatif. 2016. Improving accuracy of downscaling rainfall by combining prediction of different statistical downscale model. Journal of Water Science. 30(1):61-75.

Sari WJ. 2015. Pemodelan statistical downscaling dengan regresi kuantil komponen utama fungsional untuk prediksi curah hujan ekstrem [tesis]. Bogor (ID): Institut Pertanian Bogor.

Schroeder MA. 1990. Diagnosing and dealing with multicollinearity. Western Journal of Nursing Research. 12(2):175-187.

Vertisa D. 2014. Penentuan Bobot Optimum dengan pengganda lagrange untuk penggabungan nilai dugaan ekstrim curah hujan [skripsi]. Bogor (ID): Institut Pertanian Bogor.

Wigena AH, Djuraidah A, Mangku IW. 2014. Ensemble two return levels of generalized pareto and modified champernown distribution using linear regression. Journal of Advance and Application in Statistics. 40(2): 157-167.