Rainfall Prediction of Cimanuk Watershed Regions with Canonical Correlation Analysis (CCA)

Shailla Rustiana\textsuperscript{1}, Budi Nurani Ruchjana\textsuperscript{2}, Atje Setiawan Abdullah\textsuperscript{3}, Eddy Hermawan\textsuperscript{4}, Sinta Berliana Sipayung\textsuperscript{6}, I Gede Nyoman Mindra Jaya\textsuperscript{4,5,7}, and Krismianto\textsuperscript{7}

\textsuperscript{1} Magister Student at Department of Applied Statistics Universitas Padjadjaran and Researcher Assistant at National Institute of Aeronautics and Space (LAPAN) Bandung, Indonesia
\textsuperscript{2} Department of Mathematics, Faculty of Mathematics, and Natural Sciences Universitas Padjadjaran, Indonesia
\textsuperscript{3} Department of Computer Science, Faculty of Mathematics, and Natural Sciences Universitas Padjadjaran, Indonesia
\textsuperscript{4,5,7} Researcher at Center of Atmospheric Science and Technology Center National Institute of Aeronautics and Space (LAPAN)
\textsuperscript{6} Department of Statistics, Faculty of Mathematics, and Natural Sciences Universitas Padjadjaran, Indonesia

E-mail: shailla.rustiana@gmail.com

Abstract. Rainfall prediction in Indonesia is very influential on various development sectors, such as agriculture, fisheries, water resources, industry, and other sectors. The inaccurate predictions can lead to negative effects. Cimanuk watershed is one of the main pillar of water resources in West Java. This watershed divided into three parts, which is a headwater of Cimanuk sub-watershed, Middle of Cimanuk sub-watershed and downstream of Cimanuk sub-watershed. The flow of this watershed will flow through the Jatigede reservoir and will supply water to the north-coast area in the next few years. So, the reliable model of rainfall prediction is very needed in this watershed. Rainfall prediction conducted with Canonical Correlation Analysis (CCA) method using Climate Predictability Tool (CPT) software. The prediction is every 3 months on 2016 (after January) based on Climate Hazards group Infrared Precipitation with Stations (CHIRPS) data over West Java. Predictors used in CPT were the monthly data index of Nino3.4, Dipole Mode (DMI), and Monsoon Index (AUSMI-ISMI-WNPMI-WYMI) with initial condition January. The initial condition is chosen by the last data update. While, the predictant were monthly rainfall data CHIRPS region of West Java. The results of prediction rainfall showed by skill map from Pearson Correlation. High correlation of skill map are on MAM (Mar-Apr-May), AMJ (Apr-May-Jun), and JJA (Jun-Jul-Aug) which means the model is reliable to forecast rainfall distribution over Cimanuk watersheds region (over West Java) on those seasons. CCA score over those season prediction mostly over 0.7. The accuracy of the model CPT also indicated by the Relative Operating Characteristic (ROC) curve of the results of Pearson correlation 3 representative point of sub-watershed (Sumedang, Majalengka, and Cirebon), were mostly located in the top line of non-skill, and evidenced by the same of rainfall patterns between observation and forecast. So, the model of CPT with CCA method is reliable to use.
1. Introduction
Rainfall prediction in Indonesia is very influential on various development sectors. The inaccurate of predictions can lead to negative effects. Indonesia’s rainfall has a very large variability. Besides the influence of climate phenomena such as El Nino Southern Oscillation (ENSO) on the Pacific Ocean, Indian Ocean Dipole (IOD) on the Indian Ocean, and the Monsoon (Asian and Australian) causing extreme climate impacts such as the dry and rainy season that is longer than normal conditions [4]. It is certainly affects to various of life sectors in Indonesia. Java Island is the island with the biggest population in Indonesia and make various development sectors center on there. Java is also the largest island affected by climate phenomena compared to other islands in Indonesia, because Java is the center of the Asian-Australian monsoon region [10].

In the climate forecast system the result of research by [9], the significant common power and coherence between ENSO IOD, ENSO Indian Summer Monsoon Rainfall (ISMR) and IODISMR have clearly brought out the dominant time-period when they interact with each-other. Based on this theory, it needed a fairly accurate prediction method for West Java Province’s rainfall, especially on the area of Cimanuk Watershed, because this watershed is the one of the main support of water resources in West Java who can affect agricultural sector [11].

From the various prediction methods commonly used, known a method name as Canonical Correlation Analysis (CCA). In this study, the CCA method was run using Climate Predictability Tool (CPT) software. Climate prediction model results from CPT output is expected will give a fairly accurate prediction for West Javas rainfall.

The observation region in this research were three representative points of Cimanuk Watershed in West Java, which is divided into three parts sub-watershed (Upper, Middle, and Downstream), as shown in Figure 1.

![Figure 1. Map of observation region on Cimanuk Watershed](image)
The data used on this research were monthly data from January 2001-2016 (for predictors, as initial condition) consisting of: Nino3.4 index obtained from National Oceanic and Atmospheric Administration (NOAA); Dipole Mode Index (DMI) obtained from the website of the Japan Agency for Marine-Earth Science and Technology (JAMSTEC); Monsoon Index the calculation result of U wind data, from NOAA; and rainfall data of Climate Hazard Group Infrared Precipitation with Station (CHIRPS) over western part of Java Island from January 2001 - December 2015 as predictant.

As already reported by [3], found the lag time of predictor (SST Nino 3.4) is 2 to 3 months later. While it take only one month when we used IOD. We suspect it related well to the location of SST. The SST Nino 3.4 is located a little bit far away from Indonesian Maritime Continent comparing with IOD is located near from Indonesian Maritime Continent. While the Monsoons data dont have lag time because their pattern were sinusoidal. The predictant data we used was CHIRPS rainfall data, because the data was satellite rainfall data reanalysis with best resolution, ie 0.05°×0.05° or 50 km [2].

The tool we used were the set of computers with Microsoft Office software, Opening Grid Analysis and Display System (OpenGrADS) version 2.02 could be downloaded at (http://opengrads.org) and the Climate Predictability Tool (CPT) version 14 could be downloaded at (http://iridl.columbia.edu/). The canonical correlation coefficient values between a pair of canonical variables is based on a linear relationship. If the variables are not linearly related, then the relationship can not be explained by the canonical correlation coefficient. Canonical correlation analysis maximized the linear relationship between the set of variables [1].

The canonical correlation equation is:

$$
Corr(U, V) = \frac{Corr(a^T x, b^T y) = \frac{Cov(U, V)}{\sqrt{Var(U)} \sqrt{Var(V)}} = \sqrt{a^T S_{xy} \sqrt{b^T S_{yy} b}}}{(1)}
$$

with:

- **U**: Predictor = a1Ty + a2Ty + a3Ty + a4Ty + a5Ty + a6Ty
  (Nino3.4 + DMI + Monsoon Index: AU5MI, 8SMI, ISMI, WNlPMI, WY5MI)

- **V**: Predictant = b1T x + b2T x + ...bnT x
  (Grided rainfall of West Java)

- **S**: Matriks X,Y

Procedure of this study consisting of 3 step, as shown in Figure 2 below:

**2. Main Result**

Climate Predictability Tool (CPT) is a software based on windows that was used to develop seasonal climate prediction models based on Model Output Statistics (MOS) [8]. The result for spatial analysis of rainfall prediction map (Fig. 3) shows that the average values of 3 monthly rainfall forecast range between 0-1500 mm/3month. Red colours of the map means dry season, and the Blue colours means rainy season. Forecast rainfall on FMA (Feb-Mar-Apr) till MAM (Mar-Apr-May) were the rainy season, and MJJ (May-Jun-Jul) till JJA (Jun-Jul-Aug) season was the dry season, while the AMJ (Apr-May-Jun) and JAS (Jul-Aug-Sep) were the transition season.

The results of Pearson Correlation to explain the skill map of model or the correlation between the predictor and predictant of the model. Pearson correlation values range between -1 to 1. Where the Pearson correlation is the correlation between x variables (predictor) and y variables (predictant) that affect the results of prediction. Pearson correlation directly proportional to
Figure 2. Procedure Research of CCA

Figure 3. Forecast rainfall map of West Java

the ROC curve. The results in fig.4 showed that Pearson correlation value is mostly positive high (0.3-0.9) for every season on the area of observation Cimanuk sub-watersheds.

Figure 4. Pearson correlation as skill map of CPT model.
The good prediction model from CPT output indicated by coefficient of CCA (CCA score), shown on Table 1 by the result of formula on equation (1). CCA is a method to find the best linear combination between two multivariate datasets that maximize the correlation coefficient between them. The CCA only focuses on the correlation between a linear combination of the variables in one set and linear combination of the variables in another set. The idea is first to determine the pair of linear combinations having the largest correlation. Next, determine the pair of linear combinations having largest correlation among all pairs uncorrelated with the initially selected pair, and so on. The pairs of linear combinations are called the canonical’s variables, and their correlations are called canonical correlation [6]. This is particularly useful to determine the relationship between criterion measures and the set of reviews their explanatory factors. Involves this technique, first, the reduction of the dimensions of the two multivariate datasets by projection, and second, the calculation of the relationship (measured by the correlation coefficient) between the two projections of the datasets [7]. The basic idea behind forecasting with CCA is straightforward: simple linear regressions are constructed that relate the predictant canonical variables (x) to the predictor canonical variables (y). In this case the entire (I1) field x(t) is used to forecast the (J1) field y(t+), where  is the time lag between the two fields in the training data, which becomes the forecast lead time [12]. In applications with atmospheric data it is typical that there are too few observations n relative to the dimensions I and J of the fields for stable sample estimates (which are especially important for out-of-sample forecasting) to be calculated, even if n ≥ max (I, J) so that the calculations can be performed. It is therefore usual for both the x and y fields to be represented by series of separately calculated truncated principal components [12] Based on the equation (1), the canonical correlation coefficient (CCA Score) is the correlation between this pair of canonical variables, X (V) and Y (U), and is the largest possible correlation between pairs of linear combinations of these two data sets (x and y variables).

| Season | CCA Mode |
|--------|----------|
|        | 1  | 2  | 3  | 4  | 5  | 6  |
| FMA    | 0.94 | 0.85 | 0.39 | 0.87 | 0.66 | 0.30 |
| MAM    | 0.88 | 0.76 | 0.49 | 0.91 | 0.70 | 0.24 |
| AMJ    | 0.77 | 0.57 | 1.00 | 0.95 | 0.83 | 0.04 |
| MJJ    | 0.91 | 0.73 | 0.72 | 0.48 | 0.80 | 0.15 |
| JJA    | 0.68 | 0.61 | 1.00 | 0.79 | 0.66 | 0.01 |
| JAS    | 0.76 | 0.69 | 0.37 | 0.18 | 0.35 | 0.10 |

The CCA score per 3 month of the observation season are more than 0.7 average, which means the model was accurate enough to estimate rainfall on the observation region. Time series analysis of Observation rainfall (red line) and Hindcast rainfall from CPT (green line) for 3 representative point of sub-watersheds (Cirebon, Majalengka, and Sumedang) on FMA season (Fig. 5) were close together, and the forecast rainfall for FMA 2016 (cross point) on the normal layer (green layer). On JAS season (Fig. 6), time series analysis between observation and hindcast also close together. That means the model accurate enough to use. However the forecast rainfall for JAS 2016 on the above normal layer (red layer).

To determine the reliability of the output CPT models was carried out by analyzed the Relative Operating Characteristic (ROC) curve. ROC curve is a reliability level of the models. Linear straight lines on the ROC curve is a non-skill line. Curve over non-skill line is a positive skill (reliable), while the curve under non-skill line is a negative skill (unreliable) [5].

The ROC curve for 3 representative point of sub-watersheds (Cirebon, Majalengka, and
Figure 5. As same as Fig. 3, but for 6 months later.

Figure 6. The time-series of the Jakarta monthly rainfall intensities for period January 1995 to December 2008.

Sumedang) on FMA season (Fig. 5) were on the top of non-skill linear line, for above normal curve (red line) and below normal curve (blue line). The ROC curve for on JAS season (Fig. 6) were also on the top of non-skill linear line, for above normal curve and below normal curve. These ROC curve were proportional with Pearson Correlation. So, that means the model predictions from CPT with CCA method were reliable to use.

3. Conclusion
Rainfall prediction with CCA method of CPT software was accurate enough to estimate West Java’s rainfall on the area of Cimanuk Watersheds, by the results analysis of Canonical Correlation Analysis (CCA) score. The CCA score per 3month of the observation season were more than 0.7 average, which means the model was accurate enough. Time series analysis of Observation rainfall (red line) and Hindcast rainfall from CPT (green line) for 3 representative points of Cimanuk sub-watersheds (Cirebon, Majalengka, dan Sumedang) on FMA and JAS season were close together. That means the model was also accurate enough. ROC curve at the 3 representative point of Cimanuk sub-watersheds on FMA and JAS season were above the linear line of non-skill, for above normal curve (red line) and below normal curve (blue line). It was clearly that the CPT model was reliable or good enough to used for predict West Javas rainfall.

Acknowledgment
The work reported in this paper is supported by the Directorate General of Higher Education, Ministry of Research-Technology and Higher Education, Indonesia, through Penelitian Kerjasama Luar Negeri (KLN) dan Publikasi Internasional 2016.
References

[1] Asbah M F, Sudarno, and Diah S 2013 Penentuan Koefisien Korelasi Kanonik dan Interpretasi Fungsi Kanonik Multivariat *Jurnal Gaussian* 2(2) pp 119–128

[2] Funk C, Pete J, Martin F L, Diego H P, James P V, James D R, Bo E R, Gregory J H, Joel C M, and Andrew P V 2014 A Quasi-global Precipitation Time Series for Drought Monitoring *USGS, EROS Data Center, Sioux Falls, SD*

[3] Hermawan E 2012 Model Interkoneksi Kejadian EL Nino dan Dipole Mode Sebagai indikasi Awal Datangnya Musim Kering/Basah Panjang di Kawasan Barat Indonesia *Orasi Ilmiah Pengukuhan Profesor Riset Bidang Meteorologi*

[4] Jourdain C, Alexander S G, Andrea C U, Aurel F M, and Karumuri A 2013 The Indo-Australian monsoon and its relationship to ENSO and IOD in reanalysis data the CMIP3/CMIP5 simulations *Clim Dyn* 41 pp 3073 –3102, DOI 10.1007/s00382-013-1676-1

[5] Kadarsah 2010 Aplikasi ROC untuk uji kehandalan model HYBMG *Jurnal Meteorologi dan Geofisika* 11(1) pp 32–42

[6] Johnson R A and Dean W W 2007 Applied Multivariate Statistical Analysis 6th Edition *United States of America: Pearson Prentice Hall*

[7] Malacarne R L 2014 Canonical Correlation Analysis *The Mathematica Journal* 16 pp 1–22

[8] Mason S J Seasonal forecasting using the Climate Predictability Tool (CPT) *Science and technology infusion Climate Bulletin* In: Mason S.J. editor. 36th NOAA Annual Climate Diagnostics and Prediction Workshop [Internet], 2011, [downloaded April 1, 2014]

[9] Pokhrel S, Chaudhari H S, Subodh K S, Ashish D, Yadav R K, Kiran S, Mahaparta S, and Suryachandra A R 2012 ENSO, IOD, and Indian Summer Monsoon in NCEP climate forecast system *Clim Dyn* 39 pp 2143–2165, DOI 10.1007/s00382-012-1349-5

[10] Qian J H, Robertson A W, and Moron V 2010 Interaction Among ENSO, the Monsoon, and Diurnal Cycle in Rainfall Variability Over Java, Indonesia *Journal of the Atmospheric Sciences* 67 pp 3509 – 3524

[11] Susetyaningsih A 2012 Pengaturan penggunaan lahan di daerah hulu DAS Cimanuk sebagai upaya optimalisasi pemanfaatan sumberdaya air *Jurnal Konstruksi* 10(01) ISSN: 2302-7312

[12] Wilks D S 2006 Statistical Methods in the Atmospheric Sciences Second Edition *London, United Kingdom: Elsevier Academic Press Publications*