Statistical downscaling to predict drought events using high resolution satellite based geopotential data

H Kuswanto1*, I L Yuliatin1 and H A Khoiri1

1Department of Statistics, Faculty of Mathematics, Computing and Data Science-Institut Teknologi Sepuluh Nopember (ITS) Indonesia

*Corresponding Author: heri_k@statistika.its.ac.id

Abstract. Drought prediction is a very challenging work due to high degree of uncertainty in the climate system. Geopotential height has been investigated as one of the dominant variables that can be used to predict drought events. This paper discussed the use of high resolution satellite based (reanalysis) data as the predictor of drought events, resulting on a high dimensional dataset. To deal with this, dimension reduction has been carried out by using Principle Component Analysis (PCA), prior to the development of the downscaling models which incorporate the past SPI (Standardized Precipitation Index) combined with the geopotential height at some specific atmospheric levels i.e. 500hPa, 850hPa, 900hPa, 975 hPa and 1000hPa. The SPI, as the drought risk measure is derived from the reduced dimension of precipitation data observed from the corresponding meteorological stations, while the geopotential height is reduced from gridded high resolution data. The downscaling process found the best model to predict the drought risk with various degree of R-squares. The outsample validation showed that predicting drought using SPI3 (three month period SPI) with geopotential at the 900hPa level as the predictor outperforms the others with R-square reaching 77%.

1. Introduction

Drought is a natural disaster which leads to a cross sectoral impact. Nusa Tenggara Timur (NTT) is one of the most vulnerable regions in Indonesia toward drought [1]. The statistic of Disaster Management Office (BPBD) Indonesia stated that there were 20 districts suffered from severe drought in 2015, impacted of 270 villages [2]. According to the Food Security Monitoring Bulletin, 10 districts in NTT Province have been the first priority districts which have been seriously impacted by drought. The meteorological office (BMKG) Indonesia has been effectively monitor the weather condition and tried to accurately describe the rainfall condition in order to predict the future drought events, however the prediction cannot be made accurately due to high variability of the topography in particular of regions located near the sea. To deal with this condition, statistical downscaling technique needs to be applied by developing models which are able to characterize the drought risk.

Basically, statistical downscaling uses the relationship between atmospheric circulation with precipitation to predict the changes on the local scale [3]. The local precipitation information is obtained from downscaling technique by transforming the global scale information into a local scale [4]. The statistical downscaling technique has been applied by many researchers to predict some weather phenomena (see Busuioc et al. [5], Uvo et al. [6], Fernandes [7] among others). This paper focused on predicting drought risk in NTT Indonesia. In this case, drought risk is represented by the Standardized Precipitation Index (SPI). There are some advantages of using SPI such as the ability to be calculated....
on different scale of time; SPI serves as an alert to drought events; SPI provides information about the severity level of the events, and SPI is simpler than the Palmer Drought Severity Index (PDSI) developed by Palmer [8]. The SPI identification involves of calculating the Joint Probability Function [9]. Furthermore, the dowscaling process needs to use predictors in the model. This paper investigated the role of geopotential height satellite data as the indicator of drought events. Ribeiro and Pires [4] used geopotential height as a potential predictor to build SPI prediction model in Portugal. Another works have used geopotential height as the predictor of weather events e.g. Huth and Kysely [10], Sailor et al. [12] as well as Chen and Chen [13].

The geopotential height is an atmospheric variable where the dataset is available on grid point basis. This research used high resolution grid scale i.e. 0.25° x 0.25°, leading to high dimensional dataset. Therefore, pre-processing needs to be carried out prior to applying the downscaling process. In this case, the Principle Component Analysis (PCA) is applied to reduce the dimension of the dataset. This work investigated the satellite based geopotential height at a certain atmospheric level which can be a leading indicator to predict the drought risk in NTT. Moreover, the downscaling process used dynamic regression models involving the time lags of the corresponding variable as the predictors due to presence of temporal correlation. The inclusion of time lags has never been applied in the works mentioned above.

2. Materials and Method
Statistical downscaling is a transfer function describing the functional relationship of global atmospheric circulation with local climate variables. The aim of statistical downscaling is to find the relationship between global and local scale of the climate parameters in order to predict the local scale climate variables onset, past or in the future. In the downscaling process, the gridded variables on the global scale are the predictors and the gridded variables on the local scale as the response [14].

This research used two different dataset i.e. monthly observed rainfall data recorded at the meteorological stations and Era-Interim reanalysis dataset (the geopotential height at 500, 850, 875, 900 and 1000 hPa) with grid resolution of 0.25° x 0.25° covering of 8°S-12°S and 118°E-125°E. Both dataset span from 1999-2015.

The steps of analysis are described as follows:
1. Pre-process the data using Principal Component Analysis (PCA), both the independent and dependent variables.
2. Calculate the SPI from the pre-processed rainfall data at the following scales: 1 month (SPI1), 2 month (SPI2) and 3 month (SPI3).
3. Split the data into two parts i.e. insample data covering the periods of 1999 to 2014, and testing data which covers the last 1 year observations.
4. Carry out statistical downscaling by building regression models involving the lag of SPI and geopotential height at different levels as the predictors.
5. Validate the model for outsample data using coefficients of determination (R-squares).

3. Results And Discussion

3.1. Stylized facts about rainfall data
The rainfall dataset has been collected from 9 meteorological stations covering the whole NTT province. The presented statistic in this section is the result of pre-processed data involving of imputation for missing values and outlier detection. Figure 1 depicts the monthly rainfall average from 1999 to 2015 observed at 9 stations in NTT.

From the figure 1, we observe that the rainfall observed at all meteorological stations showed similar pattern i.e. monsoon. The precipitation in Frans is relatively higher than in other places. The periods of May to October seems to be dry periods with very low rainfall intensity, while December to February have higher precipitation rate. Nevertheless, the precipitation rate is significantly below the average in Indonesia.
Dimension Reduction of Rainfall Data Using Principle Component Analysis (PCA)

The aim of the dimension reduction is to find a response variable which can preserve the data variability without losing many information in other variables. The number of component with the corresponding eigen value can be seen is listed in Table 1. Having more than one PC does not increase the explained information significantly. With one factor, the principle component will be able to explain 68.4% total variation of the data.

Table 1. Eigen value and total variance explained.

| Component | Eigen Value | Proportion of Variance |
|-----------|-------------|------------------------|
| 1         | 6.1557      | 0.684                  |
| 2         | 0.6683      | 0.074                  |
| 3         | 0.5855      | 0.065                  |
| 4         | 0.4841      | 0.054                  |
| 5         | 0.3422      | 0.038                  |
| 6         | 0.2925      | 0.032                  |
| 7         | 0.2396      | 0.027                  |
| 8         | 0.1864      | 0.021                  |
| 9         | 0.0458      | 0.005                  |

3.2. Standardized Precipitation Index (SPI)

The SPI is obtained from the reduced dimension of rainfall data. Table 2 listed the characteristic of SPI (for three kinds of SPI calculation) based on the drought intensity. We can see clearly that the number of dry events dominates the climate condition, in particular of using the SPI1 and SPI2 indicators. Moreover, based on Figure 2 we see that those three different SPI gave similar pattern.
### Table 2. Drought intensity in NTT.

| Criteria       | Number of event |
|----------------|-----------------|
|                | SPI1 | SPI2 | SPI3 |
| Severe drought | 5    | 6    | 6    |
| Medium drought | 9    | 7    | 7    |
| Drought        | 21   | 19   | 18   |
| Wet            | 16   | 19   | 21   |
| Medium wet     | 7    | 6    | 5    |
| Severe wet     | 5    | 4    | 5    |
| Total dry event| 35   | 32   | 31   |
| Total wet event| 28   | 29   | 31   |

### Figure 2. Annual average of SPI

The SPI analysis showed also that drought events mostly happened in 2004, while 2010 tends to be wet.

#### 3.3. Dimension Reduction of Satellite Geopotential Data using Principle Component Analysis (PCA)

The geopotential height data analyzed in this paper are observed at six different atmospheric levels i.e. 500 hPa, 850 hPa, 875 hPa, 900 hPa and 100 hPa, where all of them have the same period as rainfall dataset. Using grid scale of 0.25 x 0.25, we have in total 493 observations at each levels. The scree plots indicated that only one component has eigen value greater than one. It indicates that the data variation can be explained by 1 PC for each level. The scatter plots show that most of the cases have a linear pattern, which means also that we can proceed the statistical downscaling by linear regression models. Several plots show relatively random pattern indicating low relation between SPI and geopotential. The statistical downscaling process is carried out by building regression models between geopotential data as the predictor variable and SPI as the response. The multiple linear regression models listed in Table 3 have fulfilled the required assumptions such as normally distributed residual terms and no multicollinearity among predictors. The idea of involving SPI lags in the models is to have dynamic models with capacity to predict future drought event utilizing past information. All models are performed using in-sample data.
Table 3. Regression models for the downscaling process.

| Level  | Regression equation |
|--------|---------------------|
| 500 hPa | SPI1 = -63.1 + 0.226 SPI1_{t-1} - 0.000182 Z_{500} + 0.000132 Z_{500} + 0.0000163 Z_{500} - 0.000122 Z_{500} + 0.000075 Z_{500} + 0.000087 Z_{500} - 0.000061 Z_{500} |
| 850 hPa | SPI1 = 31.4 + 0.210 SPI1_{t-1} + 0.142 SPI1_{t-2} - 0.000197 Z_{850} + 0.000118 Z_{850} - 0.000126 Z_{850} - 0.000144 Z_{850} |
| 700 hPa | SPI1 = 20.1 + 0.841 SPI1_{t-1} - 0.292 SPI1_{t-3} + 0.178 SPI1_{t-4} - 0.000059 Z_{850} - 0.000081 Z_{850} - 0.000084 Z_{850} |
| 900 hPa | SPI1 = 36.3 + 0.124 SPI1_{t-1} + 0.185 SPI1_{t-2} - 0.000189 Z_{900} + 0.000177 Z_{900} - 0.000155 Z_{900} |
| 875 hPa | SPI1 = 26.8 + 0.811 SPI2_{t-1} - 0.373 SPI2_{t-2} + 0.202 SPI2_{t-3} - 0.000145 Z_{900} + 0.000151 Z_{900} - 0.000129 Z_{900} |
| 975 hPa | SPI1 = 39.0 + 0.818 SPI3_{t-1} - 0.288 SPI3_{t-3} + 0.168 SPI3_{t-4} - 0.0543 Z_{975} + 0.0385 Z_{975} - 0.0518 Z_{975} |

From the models, the validation is done by comparing the observed SPI with the prediction. We perform only the comparison using geopotential at the level of 500hPa as the predictor, as shown in Figure 3 for the sake of space. From the figure, we can clearly see some mis-predictions e.g. the observation in January showed a normal condition but it is predicted as wet, moreover in September the real condition is drought but it is predicted as normal.
Figure 3. Plots of the observed and predicted geopotential height at 500 hPa for: (a) SPI1, (b) SPI2, (c) SPI3.

Overall, from the plots we observed that prediction with SPI3 leads the lowest error. In order to evaluate the performance, two criterias are used i.e. mean square error (MSE) and R-square. Table 4 listed both criterias for insample and outsample data.

Table 4. MSEs and R-squares for insample and outsample data.

| Level     | Insample MSE | Insample R-Sq | Outsample MSE | Outsample R-Sq |
|-----------|--------------|---------------|---------------|----------------|
| 500hPa    | 0.9214       | 11.30%        | 0.68037       | 10.70%         |
|           | 0.489        | 53.60%        | 0.764154      | 69.07%         |
|           | 0.403        | 61.10%        | 0.47751       | 61.42%         |
| 850hPa    | 0.852        | 18.60%        | 0.460085      | 32.06%         |
|           | 0.463        | 56.40%        | 0.81037       | 69.31%         |
|           | 0.395        | 62.10%        | 0.422024      | 77.22%         |
| 875hPa    | 0.9298       | 10.00%        | 0.459457      | 32.15%         |
|           | 0.498        | 52.00%        | 0.611567      | 53.68%         |
|           | 0.395        | 62.10%        | 0.589823      | 68.16%         |
| 900hPa    | 0.8463       | 19.10%        | 0.861466      | 15.19%         |
|           | 0.462        | 55.70%        | 0.602885      | 61.95%         |
|           | 0.388        | 62.70%        | 0.416284      | 77.53%         |
| 975hPa    | 0.8335       | 20.30%        | 0.986155      | 19.90%         |
|           | 0.458        | 56.20%        | 0.762638      | 56.30%         |
|           | 0.377        | 64.00%        | 0.87545       | 63.90%         |
| 1000hPa   | 0.8371       | 20.00%        | 0.931845      | 19.90%         |
|           | 0.461        | 55.80%        | 0.68828      | 56.30%         |
|           | 0.379        | 63.80%        | 0.628441      | 63.90%         |

From Table 4, using SPI1 yileds on the highest MSE and lowest R-square in all cases, while SPI3 gives the best results by increasing the R-square significantly exceeding 60%. The accuracy of the models validated on testing data can be seen in Table 5 as well. The MSE and R-square of insample and outsample data showed a consistent result, which means that the SPI3 outperforms SPI1 and SPI2. Moreover, the R-squares of SPI3 in outsample are higher than insample in some cases. Among the results of SPI3, the highest R-square is obtained by setting the geopotential height at 900 hPa level as th
predictor, with the R-square reaches about 77.53%. Furthermore, using geopotential height at 850 hPa level gives slightly different result than 900hPa, where both of them have R-squares greater than 77%.

4. Conclusion
The result of preprocessing by PCA shows that only one PC can explain more than 90% total of variance. The statistical downscaling showed that predicting drought event using three monthly SPI (SPI3) will yield on more accurate prediction. Meanwhile, using monthly SPI tends to have very low prediction capacity. The best model uses information of past SPI3 (t-1 and t-4) and geopotential at t-3 and t-4. Although it is hard to explain the rationality behind the use of those lags in the model, but the model has been statistically validated to have the best performance with the R-square exceeding 77%.

Acknowledgements
The authors acknowledge the financial support from NAS (National Academy of Science)–USAID through PEER research grant cycle 5 (NAS Sub-Grant Award Letter Agreement No. PGA-2000003422).

References
[1] BMKG, 2014. Buku informasi Peta Kekeringan dengan Metode SPI (Standardized Precipitation Index), Jakarta: BMKG.
[2] Kompas, 2015. 270 Desa/Kelurahan di 20 Kabupaten di NTT Dilanda Kekeringan. Retrieved from http://regional.kompas.com/read/2015/07/30/09274531/270.Desa.Kelurahan.di.20.Kabupaten. di.NTT.Dilanda.Kekeringan.
[3] Kasyfillah H H. 2010. Penentuan Domain Untuk Teknik Statistical Downscaling, Bogor: FMIPA IPB
[4] Ribeiro A and Pires C. 2016. Seasonal Drought Predictability In Portugal Using Statistical-Dynamical Techniques. Physics and Chemistry of the Earth, Parts A/B/C, , pp. 155-166.
[5] Busuioc A, Chen D, and Hellstrom C. 2001. Performance of Statistical Downscaling Models in GCM validation and Regional Climate Change Estimates; Application for Swedish Precipitation. *Int J. Climatology*, 21 pp.227-578
[6] Uvo C B, Olsson J, Morita O, Jinno K, Kawamura A, Nishiyama K, Koreeda N, Nakashima T. Statistical atmospheric downscaling for rainfall estimations in Kyushu Island Japan *Hydrology& Earth System Science*, 5(2) pp. 259-271.
[7] Fernandez, E. 2005. On the influence of predictors area in statistical downscaling of daily parameter. Report no. 09/2005. Norwegian Meteorological Institute, Oslo.
[8] Palmer W C. 1965. Meteorological Drought. Research Paper No. 45, U.S. Department of Commerce Weather Bureau, Washington, D.C.
[9] Betung S K P. 2015. Buku informasi Peta Kekeringan dengan Metode SPI (Standardized Precipitation Index), Tangerang Selatan: BMKG.
[10] Zorita E and Storch H V. 1998. The Analog Method as a Simple Statistical Downscaling Technique: Comparison with More Complicated Methods. *Climate*, 12 pp. 2474-2489.
[11] Huth R and Kysely J. 2000. Constructing site-specific climate change scenarios on a monthly scale using statistical downscaling. *Theoretical and Applied Climatology*, 66 pp.13-27.
[12] Sailor D J, Hu T, Li X and Rosen J N. 2000. A Neural Network Approach to Local Downscaling of GCM Output for Assessing Wind Power implications of Climate Change. *Renewable Energy*, 19 pp. 359-378.
[13] Chen D and Chen Y. 2002. Association between winter temperature in China and Upper Air Circulation Over East Asia revealed by Canonical Correlation Analysis. *Global and Planetary Change*, 780 pp.1-11.
[14] Wilhite D A. 2010. Quantification of Agricultural Drought for Effective Drought Mitigation and Preparedness: Key Issues and Challenges. Agricultural Drought Indices: Proceedings of an Expert Meeting, M. V. K. Sivakumar et al., Eds., WMO/TD1572 Geneva, WMO, pp. 13-21.