Rainfall Estimation Models Induced from Ground Station and Satellite Data

Kittisak Kerdprasop and Nittaya Kerdprasop

***Abstract*—Rainfall is an important source of water in agricultural sector of Thailand and other countries in the Southeast Asia region. Ability to estimate amount and anomaly in future rainfall is obviously great beneficial to farmers. In this paper, we propose an approach to build multivariate regression models based on the remotely sensed data obtained from the NOAA environmental satellites together with the ground-based precipitation data. The built models are for predicting and estimating amount of rainfall in a specific month. Area of study is the Nakhon Ratchasima province, which locates in the northeastern part of Thailand. We use 6-year historical data to build three models; each model is trained from the 2-year data based on three categories of rainfall in each year: drought, normal, and excessive rainfall. In this paper, we present both the models with full set of attributes and the models with selective attributes based on their importance.**

***Index Terms*—Multivariate regression, rainfall estimation, remotely sensed data, NOAA, attribute importance**

I. Introduction

THE attempt to accurately estimate amount of rainfall has appeared in a number of literature [2], [6], [10], [11], [15]. This is due to the significance of rain to agriculture and other sectors of countries such as Thailand that irrigation system is far from sufficiency and thus farmers have to rely on precipitation for their crop production.

Many studies in the past have related insufficient amount of rainfall to the drought condition of a specific country or some regions. Some researchers invented instruments and techniques for ground based estimation of rainfall [11], [12]. Several works on rainfall estimation have employed data from environmental satellite to generate rainfall estimation model [2], [6], [10], [15]. Remotely sensed data used in these research works are mostly direct precipitation monitoring and indirect interpretation based on cloud characteristics.

Recently, to characterize rainfall sufficiency or deficiency, many researchers propose the idea of using remotely sensed vegetation health or stress as substituting information to the amount of actual rainfall measured from the ground based station. Advantages of remote sensing is its low price in obtaining data, real-time accessing, and the ability to collect data in some unreachable areas. Therefore, more and more researchers are interesting in using remotely sensed data such as normalized difference vegetation index (NDVI or SMN) as a substitution for precipitation data [3], [4], [5], [14].

Besides NDVI, other indices such as vegetation health index (VHI), vegetation condition index (VCI), standardized thermal condition (SMT), and temperature condition index (TCI), generated from environmental satellite can also be used to analyze sufficient/deficient rainfall condition [1], [7], [8], [9]. We present in this paper the multivariate regression models constructed from satellite-based indices including SMN, SMT, VCI, VHI, TCI, and ground-based precipitation data. We construct three models according to the three scenarios of excessive/normal/deficient rainfall situations.

II. Data Characteristics

A. Study Area

This research used remotely sensed data in the area of Nakhon Ratchasima province in the northeastern region of Thailand. Its location [13] is shown in Fig.1

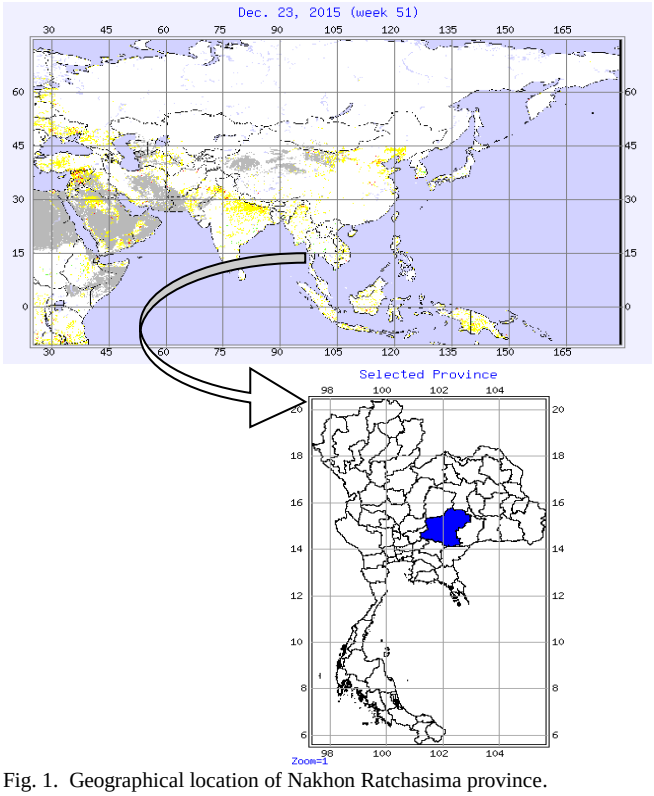

Fig. 1. Geographical location of Nakhon Ratchasima province.

Manuscript received December 30, 2015; revised January 20, 2016. This work was supported in part by the research funds from the National Research Council of Thailand and Suranaree University of Technology through the Data Engineering and Knowledge Engineering Research Units.

Kittisak Kerdprasop is the associate professor at School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: kittisakThailand@gmail.com).

Nittaya Kerdprasop is the associate professor at School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: nittaya@sut.ac.th).

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

B. *Data Source and Example*

The remotely sensed data is publicly available in the website of NOAA (National Oceanic & Atmospheric Administration), U.S.A. [13]. The NOAA operational polar orbiting environmental satellites are equipped with the advanced very high resolution radiometer (AVHRR) to detect moisture and heat that are radiated from the Earth surface. AVHRR uses several visible and infrared channels to detect radiation reflected from surface objects. Collected AVHRR data are then used to compute various indices that are helpful for crop monitoring.

We used the indices from NOAA-AVHRR available as time series data in the ASCII format. Example of NOAA-AVHRR data is shown in Fig.2. The last two columns (%Area VHI<15 and %Area VHI<35) are discarded because they are not relevant to our model analysis. The five indices shown in NOAA-AVHRR data are used in the process of model construction. Their meanings are as follows:

SMN = Smoothed and normalized difference vegetation (or smoothed NDVI) index.

It is a satellite-based index that can be used to estimate the greenness of vegetation. The value is in the range +1.0 to -1.0. Low value (0.1 or less) indicates rock or sand surface. Moderate value (0.2 to 0.5) means sparse vegetation area such as shrubs or agricultural area after harvesting. High value (0.6 to 0.9) implies dense vegetation such as forest or crops at the peak growth state.

SMT = Smoothed brightness temperature index.

It is an index that can be used to estimate the thermal condition. High SMT can lead to the dry condition of vegetation.

VCI = Vegetation condition index.

The VCI was expressed as NDVI anomaly. It can be used to estimate moisture condition. The higher VCI means the more moisture detected from vegetation. VCI less than 40 indicates moisture stress, whereas VCI higher than 60 is a favorable condition.

TCI = Temperature condition index.

TCI can be used to estimate thermal condition. TCI less than 40 indicates thermal stress, whereas TCI higher than 60 is a favorable condition.

VHI = Vegetation health index.

It is a combined estimation of moisture and thermal conditions. VHI is used to characterize vegetation health. Higher VHI implies greener vegetation. VHI less than 35 indicates moderate drought condition, whereas VHI less than 15 is severe drought. The good condition is VHI > 60.

Characteristics of these indices [13] in the area of Nakhon Ratchasima province are shown in Fig.3. The five satellite-based indices are used in our model construction, together with the precipitation data obtained from the ground based stations. Yearly precipitation information (1950-2015) in the area of study is graphically shown in Fig. 4.

year	week	SMN	SMT	VCI	TCI	VHI	%Area VHI<15	%Area VHI<35
1990	1	0.22	304.6	29.2	4.81	17.00	48.1	91.6
1990	2	0.20	305.4	29.7	4.66	17.19	46.8	91.5
1990	3	0.19	306.7	31.8	5.80	18.84	41.3	89.2

Fig. 2. Some part of NOAA-AVHRR data as time series.

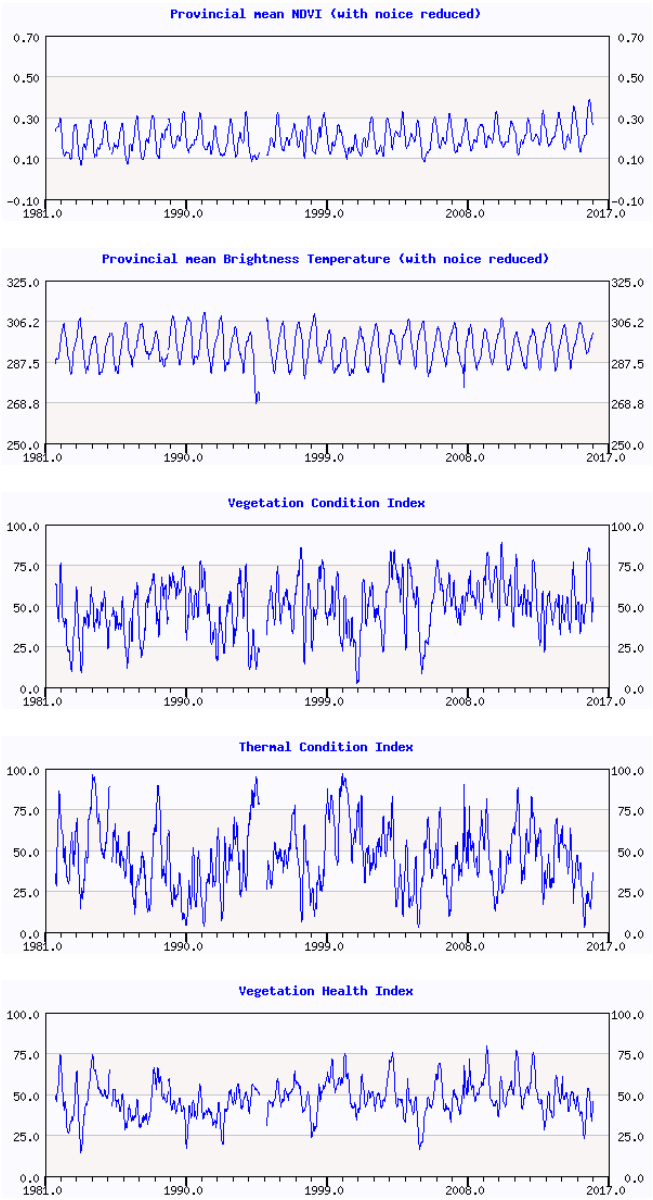

Fig. 3. The characteristics of SMN, SMT, VCI, TCI, and VHI in Nakhon Ratchasima provincial area.

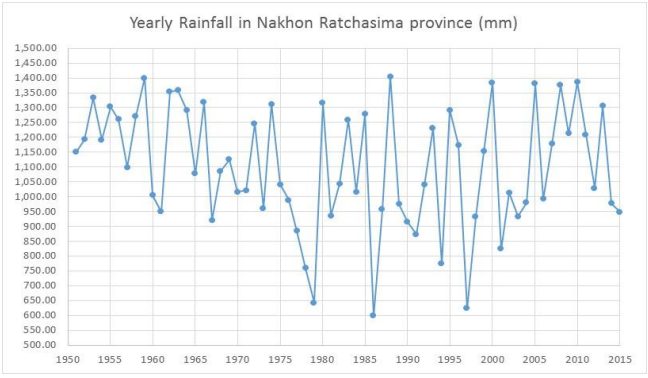

Fig. 4. The yearly rainfall (millimeters) in Nakhon Ratchasima obtained from the ground-based station (mean = 1100 mm.).

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

III. Model Construction Method

A. Data Selection and Preparation

The aim of our study is to construct a series of regression models that can be used to estimate rainfall in Nakhon Ratchasima province. The average annual rainfall in this province is around 1100 millimeters (mean value of the northeast region is 1405 mm; the average of the whole country is 1588 mm). We set 3 categories of rainfall models: drought, normal, and excessive rainfall. Distribution of monthly rainfall in each category is comparatively shown in Fig.5.

- In drought year category, we select the years:
  - 1994 (annual rainfall = 773.60 mm), and
  - 2001 (annual rainfall = 824.40 mm).

  These two years had rainfall less than 25% of the mean value.
- Normal rainfall years are:
  - 2002 (annual rainfall = 1013.50 mm), and
  - 2011 (annual rainfall = 1208.60 mm).

  The rainfall of these two years are around ±10% of the mean value.
- The excessive rainfall years are:
  - 2000 (annual rainfall = 1384.40 mm), and
  - 2010 (annual rainfall = 1386.20 mm).

  These two years had rainfall more than 25% of the mean value.

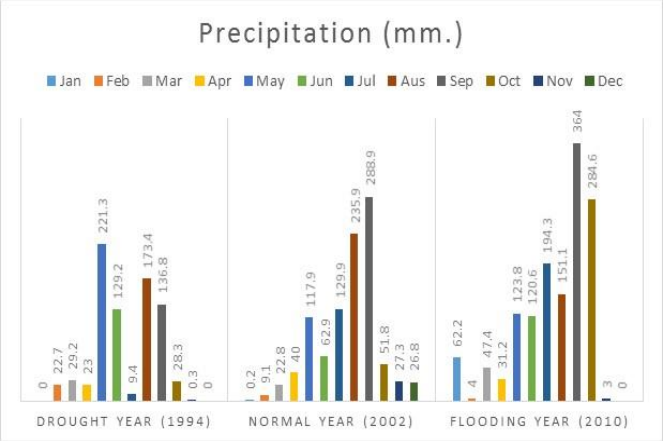

Fig. 5. Monthly precipitation patterns in Nakhon Ratchsima for each category of rainfall: drought, normal, and excessive (or flooding) years.

The satellite data in selected years are then transformed from a weekly timeframe to be monthly. Then precipitation data are added as additional column.

In this research, we have intuitive idea that satellite-based information such as VHI, SMN, and other indices that represent greenness of vegetation should be affected from precipitation in the previous months. Therefore, we create additional columns to also store satellite indices and precipitation data in the previous timeframes: one-month, two-month, and three-month before the current data instance. Example of data after the preparation process is illustrated in Fig. 6.

	month	year	SMN	SMT	VCI	TCI	VHI	Precipitation
2	1	1990	0.2002	305.9808	32.298	6.754	19.526	0.1
3	2	1990	0.170025	307.6608	44.31	20.3975	32.3525	1.2
4	3	1990	0.177375	306.7813	57.065	27.7875	42.425	24.2
5	4	1990	0.1906	304.7725	58.71	16.04	37.3725	32.8

	SMN-1	SMT-1	VCI-1	TCI-1	VHI-1	Precipitation-1
2						
3	0.2002	305.9808	32.298	6.754	19.526	0.1
4	0.170025	307.6608	44.31	20.3975	32.3525	1.2
5	0.177375	306.7813	57.065	27.7875	42.425	24.2

	SMN-2	SMT-2	VCI-2	TCI-2	VHI-2	Precipitation-2
2						
3						
4	0.2002	305.9808	32.298	6.754	19.526	0.1
5	0.170025	307.6608	44.31	20.3975	32.3525	1.2

	SMN-3	SMT-3	VCI-3	TCI-3	VHI-3	Precipitation-3
2						
3						
4						
5	0.2002	305.9808	32.298	6.754	19.526	0.1

Fig. 6. Precipitation and satellite-based indices with additional 3-month lagging period information, showing the first four months of 1990.

B. Model Construction

The data prepared from the steps explained in section *A* are then used to construct a set of regression models: one model for one category. This original models using all attributes are then analyzed for its complexity and accuracy through the R and R-square coefficients, estimation error, and importance values. We use Clementine software to analyze and plot the attribute importance.

To generate a concise model, we remove attributes that have importance value less than 0.0 from the regression models. Variable or attribute importance graphs of drought, normal, and excessive-rainfall years are illustrated in Figs.7-9, respectively.

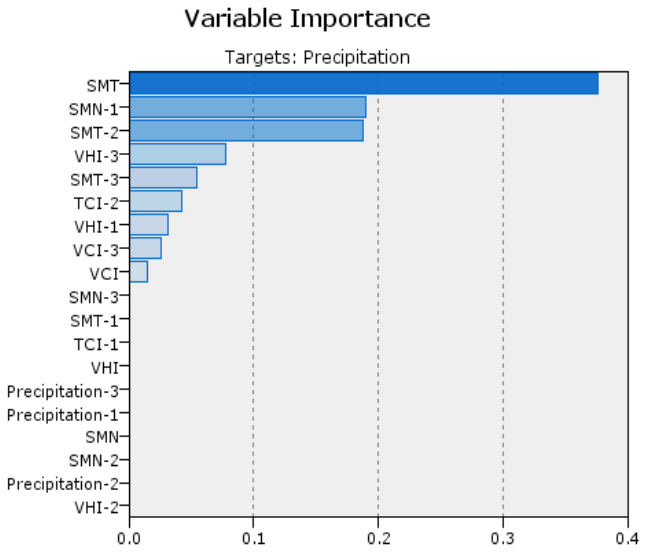

Fig. 7. Attribute importance for the drought-years: 1994 and 2001.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

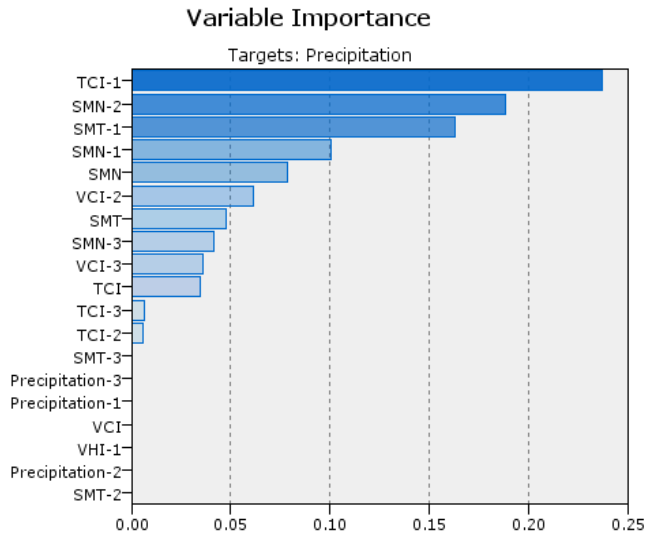

Fig. 8. Attribute importance for the normal-rainfall years: 2002 and 2011.

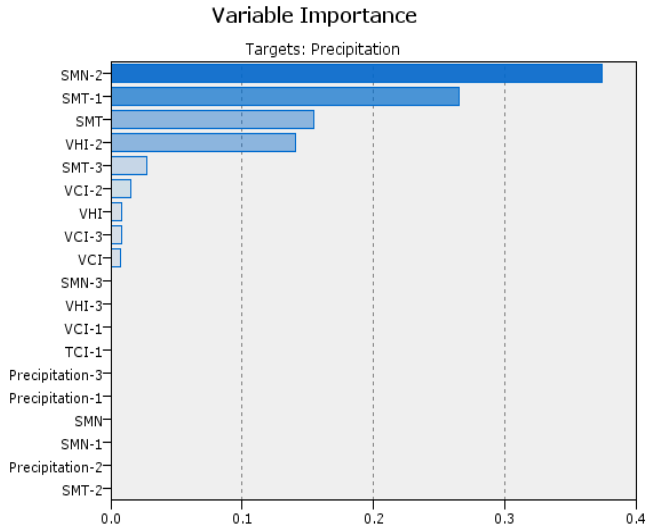

Fig. 9. Attribute importance for the excessive-rainfall years: 2000 and 2010.

IV. The Built Models

The target of our analysis is the "precipitation" attribute, which is the amount of rainfall in millimeters in the current month. The other 23 features are independent variables that are used as the predicting attributes. These attributes are SMN, SMT, VCI, TCI, VHI, SMN-1, SMT-1, VCI-1, TCI-1, VHI-1, Precipitation-1, SMN-2, SMT-2, VCI-2, TCI-2, VHI-2, Precipitation-2, SMN-3, SMT-3, VCI-3, TCI-3, VHI-3, and Precipitation-3.

The meaning of each attribute name is straightforward. For example, suppose the current time is month 4, SMN means the smoothed NDVI value of the fourth month of the year. SMN-1 means the SMN value of the third month of the same year. SMN-2 is the SMN value of the second month of the year, and SMN-3 is the value of the first month of the year. Other attribute names can be interpreted in the same manner.

The original models (the ones with all attributes) of each category of rainfall patterns have been analyzed in accordance with the ones with reduced features (which are selected from the result of importance analysis). The models of drought, normal-rainfall, and excessive-rainfall years are comparatively shown in Tables 1-3, respectively. On the construction of models with selective features, we set the regression equation to cross the origin point (y = 0). This is for the improvement of R and R-square coefficients. It is also the reason for the zero constant value in the regression models of reduced feature.

It can be seen from the results that models that are built from every attribute yield higher correlation and less error than the models constructed from the reduced features. But for this study such the full models that are constructed from more than twenty variables, it is difficult for interpretation and is also prone to over-fitting.

The over-fitting situation occurs when model shows less error when it is evaluated with the training data. But the estimation error increases when the model is used to predict other sets of information.

TABLE I
Full-feature and Reduced-feature Regression Models for Drought Years

Full-feature model	*Reduced-feature model*
Precipitation = SMN * 3440.4 + SMT * -13.98 + VCI * -1.438 + VHI * -2.7 + SMN-1 * -2702.1 + SMT-1 * 8.39 + TCI-1 * -3.793 + VHI-1 * 3.591 + Precipitation-1 * -0.8656 + SMN-2 * 2503.4 + SMT-2 * -17.48 + TCI-2 * 3.639 + VHI-2 * -6.556 + Precipitation-2 * -0.9371 + SMN-3 * 176.6 + SMT-3 * 19.47 + VCI-3 * -3.864 + VHI-3 * 10.24 + Precipitation-3 * -0.3579 + 542.0	Precipitation = SMT * -4.108 + SMN-1 * 167.3 + SMT-2 * 3.158 + VHI-3 * -1.287 + VCI * 2.065 + VHI-1 * -0.505 + TCI-2 * 3.923 + VCI-3 * 2.138 + 0.000
R = 0.974	R = 0.866
R^2 = 0.949	R^2 = 0.749
Error = 38.926	Error = 59.513

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE II
FULL-FEATURE AND REDUCED-FEATURE REGRESSION MODELS FOR NORMAL-RAINFALL YEARS

Full-feature model	*Reduced-feature model*
Precipitation = SMN * -1943.3 + SMT * -3.591 + VCI * 7.73 + TCI * -8.9 + SMN-1 * -2003.7 + SMT-1 * -21.39 + TCI-1 * -13.3 + VHI-1 * 18.75 + Precipitation-1 * -1.495 + SMN-2 * -2470.4 + SMT-2 * -14.69 + VCI-2 * 2.417 + TCI-2 * -1.605 + Precipitation-2 * -1.138 + SMN-3 * -319.4 + SMT-3 * -14.51 + VCI-3 * 3.163 + TCI-3 * -2.061 + Precipitation-3 * 0.4969 + 17451.1	Precipitation = SMT-1 * -8.735 + TCI-1 * 0.2893 + SMN-2 * 2131.5 + SMN * 199.2 + SMN-1 * -2260.7 + VCI-2 * -1.968 + SMT * 10.16 + TCI * 0.3348 + TCI-2 * 1.303 + SMN-3 * -2254.2 + VCI-3 * 3.128 + TCI-3 * -0.3742 + 0.000
R = 0.989	R = 0.945
R^2 = 0.978	R^2 = 0.892
ERROR = 33.053	ERROR = 60.307

TABLE III
FULL-FEATURE AND REDUCED-FEATURE REGRESSION MODELS FOR EXCESSIVE-RAINFALL YEARS

Full-feature model	*Reduced-feature model*
Precipitation = SMN * -4442.4 + SMT * -15.33 + VCI * 11.84 + VHI * -1.85 + SMN-1 * 10602.7 + SMT-1 * -31.89 + VCI-1 * -10.71 + TCI-1 * -4.835 + Precipitation-1 * -1.009 + SMN-2 * -9796.8 + SMT-2 * 41.02 + VCI-2 * 4.751 + VHI-2 * 14.09 + Precipitation-2 * -0.4585 + SMN-3 * 2733.8 + SMT-3 * -48.62 + VCI-3 * 2.142 + VHI-3 * -9.575 + Precipitation-3 * -0.9746 + 16380.2	Precipitation = SMT * 0.02513 + SMN-2 * -1156.3 + VHI-2 * 1.64 + VCI * 1.415 + VHI * -0.5439 + VCI-2 * -1.161 + SMT-3 * 0.7084 + VCI-3 * 1.776 + 0.000
R = 0.999	R = 0.819
R^2 = 0.998	R^2 = 0.671
ERROR = 12.845	ERROR = 110.334

To compare the performance of full feature models against the reduced feature models, we present in Fig.10 the performance comparison in terms of the number of features appeared in the model and the R coefficient value. We compare the model performance of full versus reduced features through the percentage of the decrease in number of features along with the percentage of decrease in R value.

The number of features reflects model's complexity, whereas the R-value identifies correlation of the model variables to the target of prediction. It turns out that the decreases in R-values of models with reduced features are between 4.45 to 18.02%. But we can gain much more from the decrease in terms of model complexity; the decrease is as much as 57.89%.

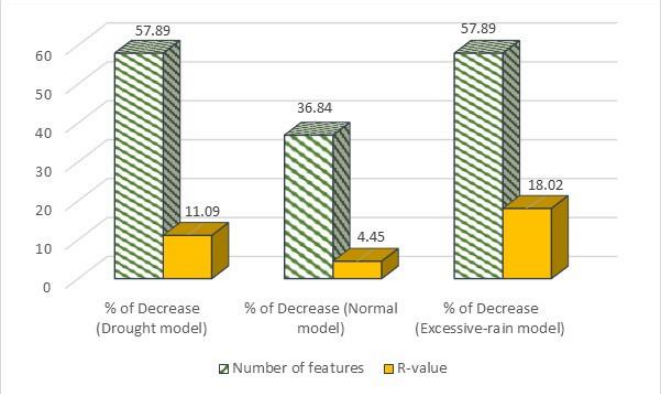

Fig. 10. Performance comparison of drought, normal, and excessive-rainfall models with full features comparative to the model with reduced features.

V. CONCLUSION

The objective of this research is the use of remotely sensed data to estimate the amount of rainfall. Advancement in remote sensing makes satellite data widely available and generated timely to facilitate rapid analysis of climate and environmental situation. We obtain climate satellite data from the NOAA-AVHRR. Index data related to vegetation monitoring have been used in this research to construct multivariate regression models for predicting amount of rainfall. Besides the satellite data, we also incorporate ground-based precipitation data for inducing an accurate models.

We construct three kinds of models: a model to estimate rainfall in a normal situation, a model for the drought year, and a model for the year with excessive rain. These models use three months lagging data to estimate the current rainfall. We firstly generate the models with full set of attributes. The obtained regression models yield very high R-square value. But the models are quite complex. We thus analyze attribute importance and select those attributes with non-zero importance value. The final results are concise models. But the trade-off is lower R-square values.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

REFERENCES

[1] Y. Bayarjargal, A. Karnieli, M. Bayasgalan, S. Khudulmur, C. Gandush, and C. Tucker, "A comparative study of NOAA-AVHRR derived drought indices using change vector analysis," *Remote Sensing of Environment*, vol. 105, 2006, pp. 9-22.

[2] T. Bellerby, K. Hsu, and S. Sorooshian, "LMODEL: A satellite precipitation methodology using cloud development modeling. Part I: Algorithm construction and calibration," *Journal of Hydrometeorology*, vol. 10, 2009, pp. 1081-1095.

[3] V. Boken, G. Hoogenboom, F. Kogan, J. Hook, D. Thomas, and K. Harrison, "Potential of using NOAA-AVHRR data for estimating irrigated area to help solve an inter-state water dispute," *International Journal of Remote Sensing*, vol. 25, no. 12, 2004, pp. 2277-2286.

[4] M. Jalili, J. Gharibshah, S. Ghavami, M. Beheshtifar, R. Farshi, "Nationwide prediction of drought conditions in Iran based on remote sensing data," *IEEE Transactions on Computers*, vol. 63, no. 1, 2014, pp. 90-101.

[5] A. Karnieli, N. Agam, R. Pinker, M. Anderson, M. Imhoff, G. Gutman, N. Panov, and A. Goldberg, "Use of NDVI and land surface temperature for drought assessment: Merits and limitations," *Journal of Climate*, vol. 23, 2010, pp. 618-633.

[6] C. Kidd, "Satellite rainfall climatology: A review," *International Journal of Climatology*, vol. 21, 2001, pp. 1041-1066.

[7] F. Kogan, "Operational space technology for global vegetation assessment," *Bulletin of American Meteorological Society*, vol. 82, no. 9, 2001, pp. 1949-1964.

[8] F. Kogan, "30-year land surface trend from AVHRR-based global vegetation health data," in F. Kogan et al. (eds.), *Use of Satellite and In-situ Data to Improve Sustainability*, 2011, pp.119-123, Springer.

[9] F. Kogan and W. Guo, "Early detection and monitoring droughts from NOAA environmental satellites," in F. Kogan et al. (eds.), *Use of Satellite and In-situ Data to Improve Sustainability*, 2011, pp.11-18, Springer.

[10] V. Levizzani, R. Amorati, and F. Meneguzzo, "A review of satellite-based rainfall estimation methods," *Technical Report MUSIC-EVK1-CT-2000-0058*, 2002, European Commission under the Fifth Framework Programme.

[11] F. Marzano, D. Cimini, P. Ciotti, and R. Ware, "Modeling and measurement of rainfall by ground-based multispectral microwave radiometry," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 5, 2005, pp. 1000-1011.

[12] F. Marzano, M. Palmacci, C. Cimini, G. Giuliani, and F. Turk, "Multivariate statistical integration of satellite infrared and microwave radiometric measurements for rainfall retrieval at the geostationary scale," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 2, 2004, pp. 1018-1032.

[13] NOAA, *STAR – Global Vegetation Health Products*, NOAA Center for Weather and Climate Prediction, Maryland, U.S.A. Available: http:// http://www.star.nesdis.noaa.gov/smcd/emb/vci/VH/index.php

[14] S. Quiring and S. Ganesh, "Evaluating the utility of the Vegetation Condition Index (VCI) for monitoring meteorological drought in Texas," *Agricultural and Forest Meteorology*, vol. 150, 2010, pp. 330-339.

[15] F. Tapiador, C. Kidd, V. Levizzani, and F. Marzano, "A neural networks-based fusion technique to estimate half-hourly rainfall estimates at 0.1º resolution from satellite passive microwave and infrared data," *Journal of Applied Meteorology*, vol. 43, 2004, pp. 576-594.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)