Big Data Analytics for Flood Information Management in Kelantan, Malaysia

[1]Aziyati Yusoff, [1]Norashidah Md Din, [2]Salman Yussof, and [3]Samee Ullah Khan

[1]College of Engineering, Universiti Tenaga Nasional, Putrajaya, Malaysia
[2]College of Information Technology, Universiti Tenaga Nasional, Putrajaya, Malaysia
[3]Department of Electrical and Computer Engineering, North Dakota State University, Fargo, United States of America
aziyati@mohr.gov.my, norashidah@uniten.edu.my, salman@uniten.edu.my, and samee.khan@ndsu.edu

***Abstract*—Big data analytics is expected to be of useful approach to many aspects of problem solving. For this research, the authors are trying to utilize this facility for flood disaster management specifically for the state of Kelantan, Malaysia. This research is considering the ordinal type data obtained from the state authorities and proposing on data manipulation through statistical inferences and big data analytics. As a result, the research is expecting for an early warning system can be developed from this study. Nonetheless, the added value approach to the analytics carried out was also considering the method of semantic network and flood ontology to design its algorithms.**

Keywords—big data analytics, flood, flood management, Kelantan flood, flood semantic network, flood information ontology.

I. INTRODUCTION

The history of data sets and storage comes a long way since the creation of computing system itself. It is the most important elements in various systems and computational structures. In fact, the need of computing and storage is due to the various information and knowledge management that people deal everyday. Hence come the age of Big Data. In this paper, the research is demonstrating on the approach of big data analytics as a problem solving for managing the flood disaster in the state of Kelantan, Malaysia.

In Malaysia, flood management is under the administration of Department of Irrigation and Drainage (DID), Ministry of Natural Resources and Environment (NRE). According to DID, every year it is estimated that 29,800 kilometers of Malaysian land is prone to flood incidence. In year 2006 and 2007, Malaysia was facing a very bad flood disaster that total loss was estimated as RM1.1 billion nationally and RM776 million for regional state assets excluding the personal loss of the affected communities nationwide [1]. However, the authority body that manage disaster occurrence in this country is the National Security Council (Majlis Keselamatan Negara, MKN). Under the Order of MKN No. 20 (Revised version), this agency has the responsibility to coordinate the management of national disaster, and to ensure that all of the national disaster policies and mechanisms are adhered and implemented at all level of disaster management units [2].

II. HYPOTHESES STATEMENTS

This research is utilizing the statistical inferences on illustrating the approach of big data analytics. This method was found to be favorable especially when dealing with Interval type large data sets [3] and has the aim of predicting the pattern of its occurrence. In this case, the study on the flood pattern and the prediction on its recursive occurrence is the pivotal point of this research.

A. Study Area

The state of Kelantan, Malaysia has the area of 117 km squared of land area bounded by latitude 6^{O}7'N, longitude 102^{O}14'E and latitude 6^{O}14'N, longitude 102^{O}9 ½ 'E and has the Kelantan River as the largest river of the state and has built its delta of partly clad with poorly drained low humid gley soils [4]. The Kelantan catchment has different soil types, but is dominated by sedentary soils on hills and mountains. The major land use of this area is agriculture (paddy, rubber and oil palm). Less than 5% of the land area is covered by developed and built land, which occurs primarily near Kota Bharu, Pasir Mas and Machang district [5].

B. Hypotheses

Each significant figures, data, information and knowledge behind this study is big data. The data might come in different forms. Under the project of Early Flood Warning System, the researchers of Universiti Tenaga Nasional (UNITEN) Malaysia are working with a number of different data types including the database query, image readings, satellite readings, and Geographical Information System (GIS) [6,7].

For this research paper, we are focusing on the Ordinal data type from perspective of statistical method. The data that will be analysed are from the rainfall and water level readings especially during the flood occurrence. It was recorded that the flood in Kelantan took place critically within the date of 20^{th} December until the first week of the new year. This happens almost every year.

Hence, we will define the hypotheses on the occurrence of the flood based on the recorded rainfall and water level data. The data was obtained from DID InfoBanjir Portal [8] with the courtesy of the department. The following statements are the hypotheses for this research.

H_0: "***Rainfall volume of more than 20mm daily which continuously pouring the state of Kelantan within 7 days will result in flood occurrence***."

H_1: "Rainfall volume of more than 20mm if and only if it started in the rural areas of Kelantan will result in flood occurrence to its urban areas especially when the river water level rises due to the fast current of the main rivers flowing from its catchment areas."

The null hypothesis, H_0, developed was upon the normal belief of local residents on the occurrence of flood incident in their area. Whilst the alternative hypotheses suggested for this study is about to be proven through statistical analysis.

For this research, we are suggesting that only when with a high density of rainfalls in rural areas will result in flood occurrence to its urban areas. This is predicted to occur within 7 days time. However, the data readings computed for the pattern study were extended within 14 days.

In fact, the readings made by DID were rigorous. The record was made every hour, and every day around the year. This makes the data a big data. For the purpose of this research we have limited our study pattern that we are considering the data with the highest read of the day on 20th December to 02 January for the last 3 years (2012, 2013, and 2014). The three consecutive years were taken as a comparison. As a record, it was reported that flood had occurred in year 2012 and 2014 but not in 2013. In the year 2012, the flood was not major. Affected areas were mostly in rural areas. The capital city of Kelantan was not threatened by the incident at all. On contrary, the year 2013 was a dry year and none were reported as flooded at all. However, the year 2014 was a major disaster to the state when it was reported that almost the whole land was about to sink. The flood has started due to the overflowing of her rivers in rural areas and had affected even to towns that were never reported as flooding before this.

Table 1 and Table 2 were constructed to illustrate this comparison. Table 1 is showing the readings for her capital city, Kota Bharu. Major flood had only occurred in year 2014. While Table 2 is showing the readings from one of the measurement station in rural areas, Laloh, Kuala Krai, Kelantan. In Laloh, it was reported to have flood in the year 2012 and 2014.

III. RESULTS AND DISCUSSIONS

The method that was used to determine the validity of the hypotheses was by Normal Distribution testing [9]. The curve pattern of the water level from its normal readings to danger level when the flood occurred is quite normal. Hence, the Z-test was carried out as part of proving the hypothesis defined.

Table 1. Rainfall Volume and Water Level Readings at Jeti Kastam station, Kota Bharu, Kelantan

Location: Jeti Kastam, Kota Bharu, Kelantan
Normal Level: 1.0 cm
Alert Level: 3.0 cm
Warning Level: 4.0 cm
Danger Level: 5.0 cm

Date	2012		2013		2014	
	Rainfall (mm)	Water Level (cm)	Rainfall (mm)	Water Level (cm)	Rainfall (mm)	Water Level (cm)
20 Dec			7.00	1.76	22.00	4.59
21 Dec			2.00	1.5	9.00	3.33
22 Dec			0.00	1.37	5.00	4.47
23 Dec	0.00	1.69	0.00	1.42	1.00	5.57
24 Dec	51.00	2.17	0.00	1.47	22.00	6.49
25 Dec	39.00	4.29	0.00	1.25	3.00	6.88
26 Dec	58.00	4.50	0.00	1.19	0.00	6.89
27 Dec	9.00	4.30	0.00	1.26	0.00	6.89
28 Dec	0.00	2.73	10.00	1.56	0.00	6.89
29 Dec	0.00	1.61	10.00	1.37	0.00	6.89
30 Dec	27.00	1.46	4.00	1.37	0.00	6.89
31 Dec	64.00	1.64	4.00	1.54	0.00	4.92
01 Jan	37.00	3.23	2.00	1.68	0.00	4.22
02 Jan	14.00	3.39	2.00	1.62	0.00	2.36
Mean	**27.18**	**2.82**	**2.93**	**1.45**	**4.43**	**5.52**
Std. Dev.	24.14	1.19	3.65	0.17	7.89	1.54
Z-Test	0.16	1.00	1.00	1.00	0.61	1.00

A. Z-Test

The Z-test was carried out to prove on the proposed hypothesis. The critical value was when the rainfall has reached 20mm, does it give the significant effect to the water level and had forced the readings to indicate on flood incidence.

First, we tested the data from the flood station of Kelantan's capital city i.e. Kota Bharu. The station was located at Jeti Kastam and Table 1 is illustrating the findings. At the station, if the water level was read less than 3.0 cm, then it is a Normal level. If the data read was more than 3.0 cm but less than 4.0 cm, it is an Alert level. If the data was between 4.0 cm to 5.0 cm, then it is a Warning level. And if the water level is 5.0 cm, it is Danger level and a sign that the city of Kota Bharu is about to sink.

By usinq the Z-test with 95% confidence level, it was found that in the year 2012, the *P-value* of the rainfall at the average of 20 mm was 0.16. Whilst, the *P-value* of the rainfall for the year 2013 and 2014 were 1.00 and 0.61 respectively. On contrary, in the year 2012 the *P-value* for the water level to reach at least the warning level showing that the flood is about to occur was 1.00. The same *P-value* returned for its next consequent 2 years.

Simultaneously, we also need to study the pattern of the rainfall and the measurement of water level at rural areas that might affect the flood incidence in Kota Bharu. Hence, from Table 2, we have found that for the year 2012, the *P-value* of the rainfall with the average of 20mm in Kuala Krai, Kelantan was 0.45. While the *P-value* of the same entity was both 1.00 for the year 2013 and 2014. However, the *P-value* for the flood to occur and reached at least its warning level was 0.00 for the year 2012, 1.00 for the year 2013 and 0.00 for the year 2014.

This means that we reject null hypothesis and the flood incidence and data readings from the rural areas do have significant effects on the flood occurrence in its urban areas.

Table 2. Rainfall Volume and Water Level Readings at Kampung Laloh,Kuala Krai, Kelantan

Location: Kampung Laloh, Kuala Krai, Kelantan
Normal Level: 17.0 cm
Alert Level: 20.0 cm
Warning Level: 22.5 cm
Danger Level: 25.0 cm

Date	**2012**		**2013**		**2014**	
	Rainfall (mm)	Water Level (cm)	Rainfall (mm)	Water Level (cm)	Rainfall (mm)	Water Level (cm)
20 Dec			3.00	18.29	4	23.45
21 Dec			0.00	17.95	15	22.71
22 Dec			0.00	17.53	41	27.75
23 Dec	2.00	26.73	0.00	17.53	0	30.47
24 Dec	78.00	31.52	1.00	17.12	0	33.71
25 Dec	33.00	35.27	1.00	17.12	0	34.17
26 Dec	16.00	35.17	0.00	17.01	0	33.98
27 Dec	5.00	31.03	0.00	16.86	0	33.98
28 Dec	18.00	26.17	34.00	17.00	0	33.98
29 Dec	0.00	25.94	35.00	17.17	0	33.98
30 Dec	14.00	25.26	0.00	17.37	0	33.98
31 Dec	37.00	29.14	0.00	17.21	0	25.98
01 Jan	12.00	31.48	1.00	17.00	0	23.51
02 Jan	14.00	29.62	1.00	16.86	0	21.06
Mean	**20.82**	**29.76**	**5.43**	**17.29**	**4.29**	**29.48**
Std. Dev.	22.18	3.53	12.35	0.42	11.32	5.16
Z-Test	0.45	0.00	1.00	1.00	1.00	0.00

B. Correlation Testing

Consequently, the study on the patterns of rainfalls and water level has led us on the need to do correlation testing. Correlation testing was created mainly to describe how closely related when there are two physical characteristics were involved. This is particularly when we made assumptions on its relationships and connections.

Figure 1, 2 and 3 showing are showing the relationship diagrams of the rainfall data with the water level readings during the 14 days of study. They are illustrating for the year 2012, 2013, and 2014 respectively for the reading station of Jeti Kastam, Kota Bharu, Kelantan.

From the diagrams, the strength of the relationship can be done by calculating its correlation coefficient. The correlation

coefficient, *r* was calculated by using the following equation [9],

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) \quad (1)$$

From formula (1), we have found out that the correlation coefficient, ***r*** for the rainfall and water level at Kota Bharu station for the period of 20 December 2012 to 2 January 2013 is equal to **0.16**. While for the year 2013/14, the r is equal to **0.41** and for the year 2014/15, the r is equal to **-0.13**. For these values of correlation coefficient, we conclude that the relationship between the rainfall and water level is very weak. Hence, there are actually other factors that have contributed to the disastrous event of flood occurrence in Kelantan.

C. Predictive Model Proposition

The coefficient correlation calculated from previous subsection was also used to find its simple linear regression. Once again, by referring Figure 1, 2 and 3 there was a straight line drawn across the plots of the chart respectively. This is the simple linear regression for the chart. This regression is meant for defining our prediction pattern for the two variables that we are computing.

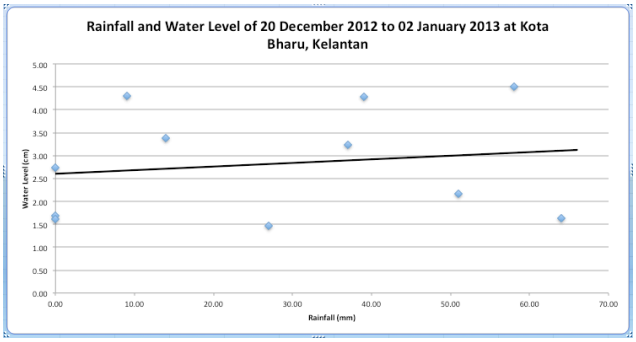

Figure 3. Rainfall and Water Level of 20 Dec 2012 to 02 Jan 2013 at Kota Bharu, Kelantan

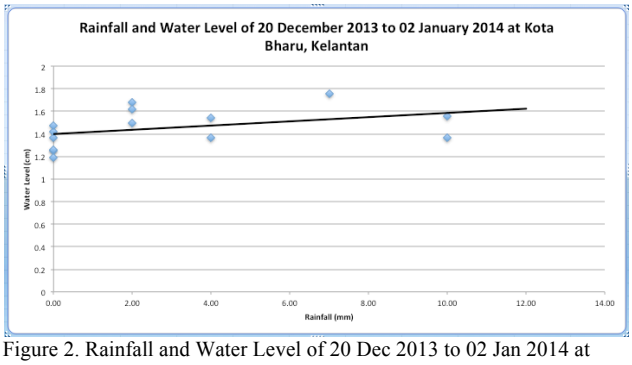

Figure 2. Rainfall and Water Level of 20 Dec 2013 to 02 Jan 2014 at Kota Bharu, Kelantan

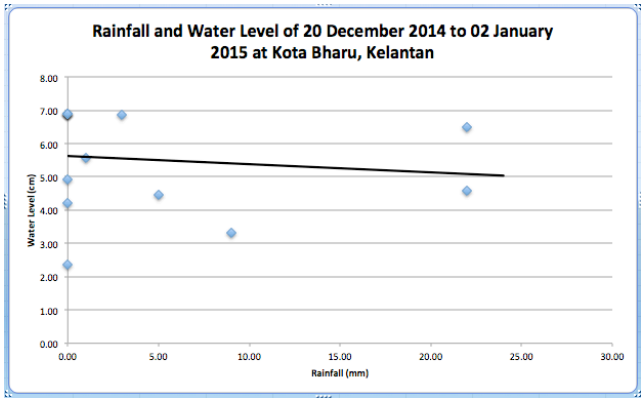

Figure 3. Rainfall and Water Level of 20 Dec 2014 to 02 Jan 2015 at Kota Bharu, Kelantan

In fact, statisticians had given that prediction interval for future observation is as the following equation [8],

$$Pred = \hat{\beta} + \hat{\beta}_1 x \pm t_{n-2,\alpha/2}\left(s\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\right) \quad (2)$$

Formula (2) is analyzing on the simple linear regression model of the relationship between the two variables and thus making prediction on its forecasting method and future outcomes. However, this model seems not to fit the flood incidence due to the event was based on specific disaster condition and very unlikely to expect that it should sustain on its performance.

Scholars had also proposing a number of practical approach in handling the weather-based disaster event for a better community management such as leveraging the weather forecasts in energy harvesting sensor systems by using cloud computing [10], Parameter ESTimator (PEST) in predictive analysis and simulation of distributed hydrological model [11] and examination of Integrated Flood Analysis System (IFAS) from the Tropical Rainfall Measuring Mission (TRMM) [12,13,14].

In fact, for this research we are considering both the readings from the rainfall data and water level as the main key parameters to determine on the alert type that should warn the public on the occurrence of the flood. Hence, to obtain such an algorithm is discussed in section IV of this paper.

IV. ONTOLOGY OF THE FLOOD INFORMATION MANAGEMENT (FIM) AND SYSTEM ALGORITHM

Creating ontology for a certain information system is a powerful tool to limit the complexity of a problem and optimize its performance [15,16,17]. The ontology is a powerful tool in translating a semantic network that we are building for a system. From this ontology it is expected that the main classes of the constructed system, its properties and relationships can be well defined.

A. FIM Ontology Design and Development

For this research, we have constructed ontology for the semantic of Flood Information Management (FIM) System for the state of Kelantan. This ontology design is expected to scrutinize the development of the system's algorithm and procedures alike. Figure 4 is illustrating the FIM and its main class and its subclasses. These classes carry specific properties to the system and their relationship between one (sub)class to another.

FIM is the main class of this system. It has a number of subclasses which one of them include the *BigDataAnalytics*, which was also represented, by another subclass of *HydrologyData*. The *HydrologyData* class was listing flood measuring stations throughout the state as its subclasses. As of year 2014, there are 13 stations that are operating in the state namely at *Aring, Dabong, GuaMusang, GunungGagau, Jeli, Jenob, KotaBharu, KualaKrai, Kusial, Laloh, PasirPutih, RantauPanjang, and Tualang*. Each station is read daily regardless of the weather or the flood condition making its data huge and requires proper management. Hence, come the need to analyze its reading by big data analytics.

B. The Algorithms

From the ontology designed in previous subsection, we will construct the algorithm on its flood prediction approach. Algorithm 1 is illustrating this prediction based on the logic readings from the classes and properties of the semantic defined. The main aim of the algorithm is to return an emergency alert procedure to its authorities and public knowledge and actions.

Algorithm 1: Flood prediction

Input: Set of rainfalls = R, water level W and P prediction of flood occurrence

Output: Emergency warning and flood alert to the authorities

Definitions: R=Set of rainfall readings, W=Set of water level readings,

```
1: Procedure Predict (W)
2: FOR R = 0.0;
3:         Do P = R++ && W++ >> AlertLevel;
4:         Return EmergencyAlert (W);
5: Else return (null);
6: End
```

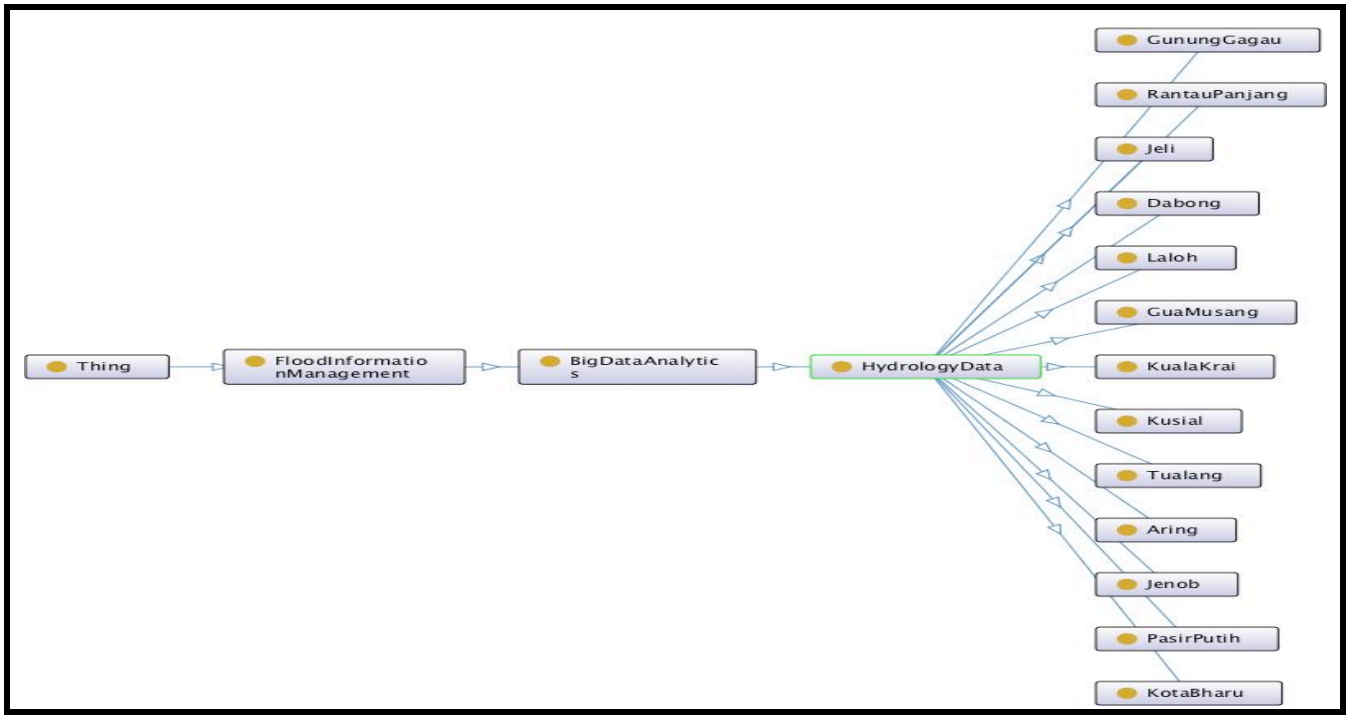

Figure 4. The Ontology of Flood Information Management for Kelantan

V. FUTURE WORKS

Many aspects for future works can be tackled from this research. Firstly, the hypotheses imposed can be further improved with other possible variables that might be taken into considerations when predicting flood occurrence. Secondly, the correlation of the measurement of water level, which triggered the flood incidence, should be tested with other contributing factors besides the rainfall. As a result, it might involve a much complex relationship of the factors and variables defined. Lastly, the semantic network and ontology relationship of this system should be further elaborated with its other supportive elements that could make the system possible to run and to produce an operating early flood warning system.

VI. CONCLUSION

This research is to study the significant of big data analytics to the flood occurrence in the state of Kelantan, Malaysia. Inferences made were on the relationship of the rainfall data with water level of the rivers. This is based on the assumptions made by public and authorities on the main factors that have affected to the flood incidence. From this study, it was found out has that the relationship between the rainfall and water level is very weak. The study suggests that it should expand its research to other contributing factors that have triggered the occurrence of the flood. Consequently, this study is also proposing on the algorithm of predicting the flood incidence. The algorithm is as a result of semantic network and ontology design. Hence, the research is expecting that an early warning alert will be produced from this research.

ACKNOWLEDGMENT

The authors wish to express the gratitude to the research team of Early Flood Warning System and Big Data Analytics of College of Engineering, UNITEN under the JICA-SATREPS Project on Research and Development for the reduction geo-hazard damage in Malaysia caused by landslide and flood and also to Big Data research team of North Dakota State University (NDSU). Heartiest appreciation goes to Public Service Department of Malaysia (JPA) as the main sponsor for this research, and Department of Irrigation and Drainage (JPS) Malaysia for providing the flood big data.

REFERENCES

[1] Pengurusan Bencana Tanggungjawab Bersama. URL: http://portalbencana.mkn.gov.my/Portal/Board/Detail?board=141&entity=7524 2014.

[2] Majlis Keselamatan Negara 2012 *Arahan No. 20 (Semakan Semula): Dasar dan Mekanisme Pengurusan Bencana Negara* (Jabatan Perdana Menteri, Malaysia), 2014.

[3] Assunçaoa M. D., Calheirosb, R. N., Bianchia, S., Nettoa, M. A., and Buyyab, R. . "Big Data Computing and Clouds: Challenges, Solutions, and Future Directions." *arXiv preprint arXiv:1312.4722* (2013).

[4] Zakaria A. S., "*The Geomorphology of Kelantan Delta (Malaysia)"* (Catena 2) pp 337-349, 1975.

[5] TCPD 2002 *Malaysia Structure Plan, Malaysia Department of Town and Country Planning* (Unpublished report TCPD: Kuala Lumpur). 2003.

[6] Yusoff A., Mustafa I.S., Yussof S., and Din N.M., "Green cloud platform for flood early detection warning system in smart city." *Information Technology: Towards New Smart World (NSITNSW), 2015 5th National Symposium on*. IEEE, 2015.

[7] Yusoff A., Din N. M., and Khan S. U., "Cloud Architecture and Big Data Analytics for Flood Management in the Landscape of Malaysia's 1Gov*Net ICT Infrastructure" in ND EPSCoR State Conference Fargo ND USA). 2015.

[8] InfoBanjir Portal, URL: http://infobanjir.water.gov.my, 2006.

[9] Navidi W., "Statistics for Engineers & Scientists (Fourth Edition)", McGraw Hill, New York, 2011.

[10] Sharma N., Gummeson J., Irwin D., and Shenoy P.. "Cloudy computing: Leveraging weather forecasts in energy harvesting sensor systems." *Sensor Mesh and Ad Hoc Communications and Networks (SECON), 2010 7th Annual IEEE Communications Society Conference on*. IEEE, 2010.

[11] Bahremand A., and De Smedt F., "Predictive analysis and simulation uncertainty of a distributed hydrological model", in Water resources management 24.12 (2010): 2869-2880.

[12] Limlahapun P., Fukui H., Yan W., and Ichinose T. "Integration of flood forecasting model with the web-based system for improving flood monitoring, and the alert system" in Control, Automation and Systems (ICCAS), 2011 11th International Conference on. IEEE, 2011.

[13] NRT Global Flood Mapping, URL: http://oas.gsfc.nasa.gov/floodmap/, 2015.

[14] Tropical Rainfall Measuring Mission, URL: http://trmm.gsfc.nasa.gov , 2015.

[15] Garrido J., and Requena I., "Towards summarizing knowledge: Brief ontologies" in Expert Systems with Applications 39.3 (2012): 3213-3222.

[16] Glimm B., Horrocks I., Motik B., Shearer R., and Stoilos G. "A novel approach to ontology classification" in Web Semantics: Science, Services and Agents on the World Wide Web, 14 (2012): 84-101.

[17] Scheuer, S., Haase D., and Meyer V. *Towards a flood risk assessment ontology–Knowledge integration into a multi-criteria risk assessment approach* in Computers, Environment and Urban Systems 37 (2013): 82-94.