IOP Conference Series: Earth and Environmental Science

Free internet datasets for streamflow modelling using SWAT in the Johor river basin, Malaysia

To cite this article: M L Tan 2014 *IOP Conf. Ser.: Earth Environ. Sci.* **18** 012193

IOP Conf. Series: Earth and Environmental Science **18** (2014) 012193 doi:10.1088/1755-1315/18/1/012193

Free internet datasets for streamflow modelling using SWAT in the Johor river basin, Malaysia

M L Tan[1]

PhD Student, T-06, Institute Geospatial Science and Technology (INSTeG), Faculty of Geoinformation and Real Estate (FGRE), Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

Email: mouleong@gmail.com

Abstract. Streamflow modelling is a mathematical computational approach that represents terrestrial hydrology cycle digitally and is used for water resources assessment. However, such modelling endeavours require a large amount of data. Generally, governmental departments produce and maintain these data sets which make it difficult to obtain this data due to bureaucratic constraints. In some countries, the availability and quality of geospatial and climate datasets remain a critical issue due to many factors such as lacking of ground station, expertise, technology, financial support and war time. To overcome this problem, this research used public domain datasets from the Internet as "input" to a streamflow model. The intention is simulate daily and monthly streamflow of the Johor River Basin in Malaysia. The model used is the Soil and Water Assessment Tool (SWAT). As input free data including a digital elevation model (DEM), land use information, soil and climate data were used. The model was validated by in-situ streamflow information obtained from Rantau Panjang station for the year 2006. The coefficient of determination and Nash-Sutcliffe efficiency were 0.35/ 0.02 for daily simulated streamflow and 0.92/ 0.21 for monthly simulated streamflow, respectively. The results show that free data can provide a better simulation at a monthly scale compared to a daily basis in a tropical region. A sensitivity analysis and calibration procedure should be conducted in order to maximize the "goodness-of-fit" between simulated and observed streamflow. The application of Internet datasets promises an acceptable performance of streamflow modelling. This research demonstrates that public domain data is suitable for streamflow modelling in a tropical river basin within acceptable accuracy.

1. Introduction

The concept of Digital Earth has originally been introduced by US Vice President Al Gore in 1998. It integrates and connects of the existing digital knowledge into the earth virtual representation. The integration of this idea into a hydrology cycle gives better visualization and understanding of movement of water between land, ocean and sky. Streamflow modelling attempts to represent the hydrology cycle digitally, from precipitation to streamflow using mathematics [1]. It is useful a tool for flood assessment, dam construction, water quality monitoring and aquatic life investigation [2].

Large quantities of data sets for different parts of the terrestrial system including land use, soil, stream network, weather condition and so on are required for modelling streamflow. However, availability and assessment of such information for the public remains a constraint [3]. Generally,

[1] To whom any correspondence should be addressed

government departments are responsible to maintain such information. The general public has difficulties to access due to bureaucratic constraints. Some government departments are imposing a processing fee on the data. Therefore, researchers or water resource managers that have no budget to purchase these produce face difficulties in streamflow modelling. Besides that, scarcity and sparse distribution of hydrology and climate monitoring stations also contribute to data collection trouble within a water resource project.

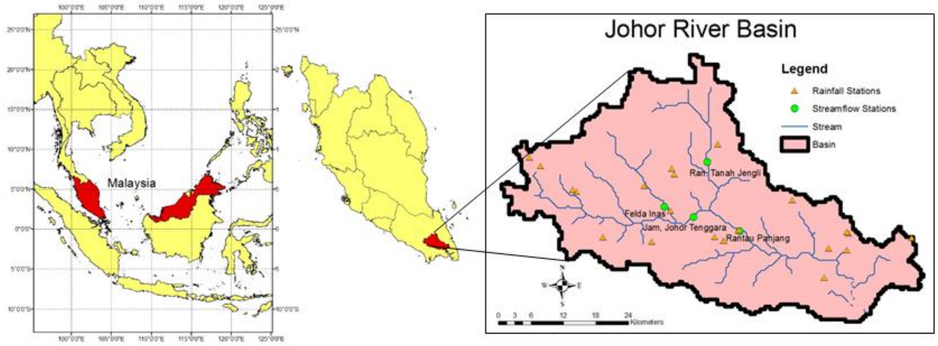
Recently, a huge amount of public domain datasets that can act as “input” to streamflow modelling is available on the Internet. These datasets are usually free of charge, easy to obtain and frequently updated compare to local sources. So, the user can perform the modelling for any basin of interest provided there is Internet access. Lacroix et al. [4] used public domain data including topography, land use, leaf area index (LAI) and climate data for a Semi-Distributed Land-Use Runoff Process (SLURP) of the Küçük Menderes Basin in Western Turkey. The result shows a good agreement of simulated and observed streamflow for the years of 1994 to 1997 with a standard error of 73%.

The Soil and Water Assessment Tool (SWAT) was chosen for this study because it is a popular streamflow model that has proven to be effective in water resource management [5,6]. The first attempt to apply public domain data sets within SWAT was conducted by Van et al. [6] for three different large river basins including the River Ganges in India, Bangladesh and Nepal; the River Blue Nile in Ethipia; and River Kagera in Rwanda, Burundi, Uganda and Tanzania. The good correlation of monthly simulated streamflow could only be obtained at River Blue Nile where NSE for Kessi and Sudan Border station is 0.8 and 0.74 respectively.

The main question is how accurate are the input parameters mainly gained from Internet sources for a tropical region. The aim of this study is to test the suitability of the application of public domain data sets as “input” parameter to generate streamflow using SWAT for Johor River Basin in Malaysia. For this goal the simulated streamflow was compared to in situ data collected by Rantau Panjang station to evaluate the effectiveness of free data sets and SWAT model.

2. Study Area

The study was conducted at the Johor River basin, located in the southern part of the Peninsular Malaysia (Fig 1). The total area of basin area is 1842.8 km^2 with an average rainfall of 2500mm. There are four streamflow monitoring stations in this area: Tanah Jengli, Felda Inas, Johor Tenggara and Rantau Panjang. In-situ streamflow was collected from Rantau Panjang station for validation process of the year 2006 and mean annual streamflow is $37.7 m^3 s^{-1}$. The basin is mostly covered by oil palm plantation, followed by rubber, tropical evergreen forest, secondary forest, water bodies and swamp forest. There are three main types of soil for this area which are Ferric (Af), Orthic (Ao) from the Acrisols group and Thionic (Jt) from the Fluvisols group. Based on Fig 2(b), the Johor River Basin is dominated by Orthic 72%, Thionic 25% and Ferric 5%. The main population area of the basin is Kota Tinggi and frequently experiences flooding events. The most severe flood event occurred in December 2006 and January 2007 which bring total loss of around USD 0.5 billion, 110 000 people were evacuated and 18 people died [7].



(a) (b) (c)

Figure 1. Study region: (a) Malaysia within the Southeast Asia; (b) Penisular Malaysia; (c) Johor River Basin.

3. Model and Data

The sections describe the model in more detail and provide an overview on the content and quality of the data used.

3.1. Soil Assessment and Tool (SWAT)

SWAT model is a physical based, continuous hydrological model which is based on concept of hydrologic response units (HRU) [5]. HRU are portions of sub-basins that represent a special soil, land use and slope combination. Surface runoff, infiltration and soil erosion processes are generated for each different HRU. There are two approaches for measuring surface runoff which are the Green & Ampt infiltration method that needs sub-daily precipitation data and the Soil Conservation Service (SCS) curve number which requires daily precipitation as selected for the modelling. Besides that, Hargreaves, Priestley Taylor and Penman-Monteith methods are options that can be selected for the evapotranspiration process in SWAT. Penman-Monteith was chosen because the climate data sets collected from Internet contain information on air temperature, solar radiation, relative humidity and wind speed. Hargreaves only requires air temperature while Priestley Taylor is similar to Penman-Monteith but does not use wind information. The main data sets that are required for SWAT modelling are land use, DEM, soil and climate data which I will further discuss in the following section.

3.2. Land use

Land use data was obtained from the U.S. Geological Survey (USGS) Land Cover Institute (LCI), i.e. Global Land Cover 2000 (GLC 2000) in http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php [8]. The GLC 2000 was produced by a series of 14 month data sets from 1998 to 2000 that monitored by the Système Pour l'Observation de la Terre (SPOT) 4 Vegetation sensor. The instrument collects images with a swath width of 2200km and a spatial resolution of about one kilometer. The GLC 2000 was available in different parts of the world. I chose the Southeast region for this study. Stibig et al. [9] evaluated the accuracy of SPOT-4 Vegetation forest cover datasets with Landsat Thematic Mapper (TM) forest classification images. The result indicates that SPOT-4 data sets corresponded or matched to 79% forest and 4% fragmented forest on TM image and the regression coefficient (R^2) is 0.96.

3.3. Soil

The Food and Agriculture Organization of the United Nations (FAO-UNESCO) provided a soil map of the world (Version 3.6) in http://www.fao.org/geonetwork/srv/en/metadata.show?id=14116 and was completed in January 2003 [10]. The soil map was prepared using the American Geographical Society of New York topographic map series at 1:5000000 scale in geographic projection with a 10km spatial resolution. Shrestha et al. [11] used the FAO soil dataset for streamflow modelling for the Bagmati River Basin at the Nepal. Their studies indicate that there is a good correlation between daily observed and simulated streamflow at Pandhera Dovan station with R^2 of 0.91 and concluded that the FAO soil data set is the best public domain soil data sets.

3.4. DEM

Movement and direction of the water flow is influenced by the topographic condition of the basin. Therefore it is inevitable to use a Digital Elevation Model (DEM). The DEM used in this study was taken from National Aeronautics and Space Administration (NASA) Shuttle Radar Topography Mission (SRTM), at 3 arc second (90m resolution) and with an accuracy of about ±16m [12]. SRTM generated elevation data for near global scale and delivery free by the USGS in http://srtm.csi.cgiar.org [13].

3.5. Climate Data

The National Centers for Environment Prediction (NCEP) Climate Forecast System Reanalysis (CFSR) developed a 31 year climate data sets (1979 to 2010) which is freely available to the public in http://rda.ucar.edu/datasets/ds093.1. [14]. Climate data sets for the study area such as precipitation, solar radiation, relative humidity, air temperature, wind speed was obtained from the CFSR distributor. The CFSR integrated two types of precipitation data sets for land area which are the CPC Merged Analysis of Precipitation (CMAP) with 2.5^o resolution and the CPC unified global daily gauge analysis data at 0.5^o resolution.

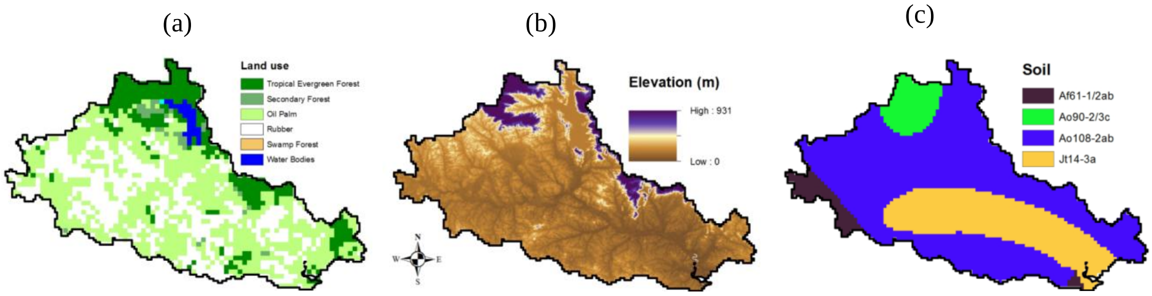


Figure 2. Input Parameters for SWAT (a) Land use; (b) Digital Elevation Model; and (c) Soil.

3.6. Model validation

The performance of the model was evaluated based on four statistical analysis parameters such as Nash-Sutcliffe Efficiency (NSE) [15]; coefficient of regression (R^2), percentage of bias (PB), root mean square error (RMSE), mean absolute error (MAE) and RMSE-observation standard deviation ration (RSR):

$$NSE = 1 - \frac{\sum_{i=1}^{n}(o-p)^2}{\sum_{i-n}^{n}(o-\bar{o})^2} \quad (1) \qquad R^2 = \left(\frac{\sum_{i=0}^{n}(o-\bar{o})(p-\bar{p})}{\left[\sum_{i=0}^{n}(o-\bar{o})^2 \sum_{i=0}^{n}(o-\bar{p})^2\right]^{0.5}}\right)^2 \quad (2) \qquad PB = \frac{\sum|o-p|}{\sum o}(100) \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(o-p)^2}{n}} \quad (4) \qquad MAE = \frac{\sum_{i=1}^{n}(o-p)}{n} \quad (5) \qquad RSR = \frac{RMSE}{\sqrt{\sum_{i=1}^{n}(o-\bar{p})^2}} \quad (6)$$

Where o and p are observed and simulated streamflow, respectively; n is the number of measured streamflow. The NSE is widely applied in hydrograph assessment in order to measure “goodness-of-fit” between simulated and observed streamflow [16]. The NSE value between 0 to 1 (ideal) is known as acceptable performance while -∞ to 0 represents unacceptable performance because of the mean observed value that is better compared to the simulated value [16,17]. R^2 can be used to evaluate the correlation between two different variables and range from -1 to 1 (perfect). The optimal value for PB is 0. A negative value shows overestimated bias while a positive value indicates an underestimation of bias of the simulated variable compared to the observed variable. RMSE is the average of differences between simulated and observed streamflow with the relationship: the smaller the value, better the accuracy. RMSE is sensitive to extreme values, so MAE also computed for the evaluation. RSR is a popular error index statistic and better modelling performance a lower RSR value. Table 1 represents the result of the validation evaluation by the different type of approaches.

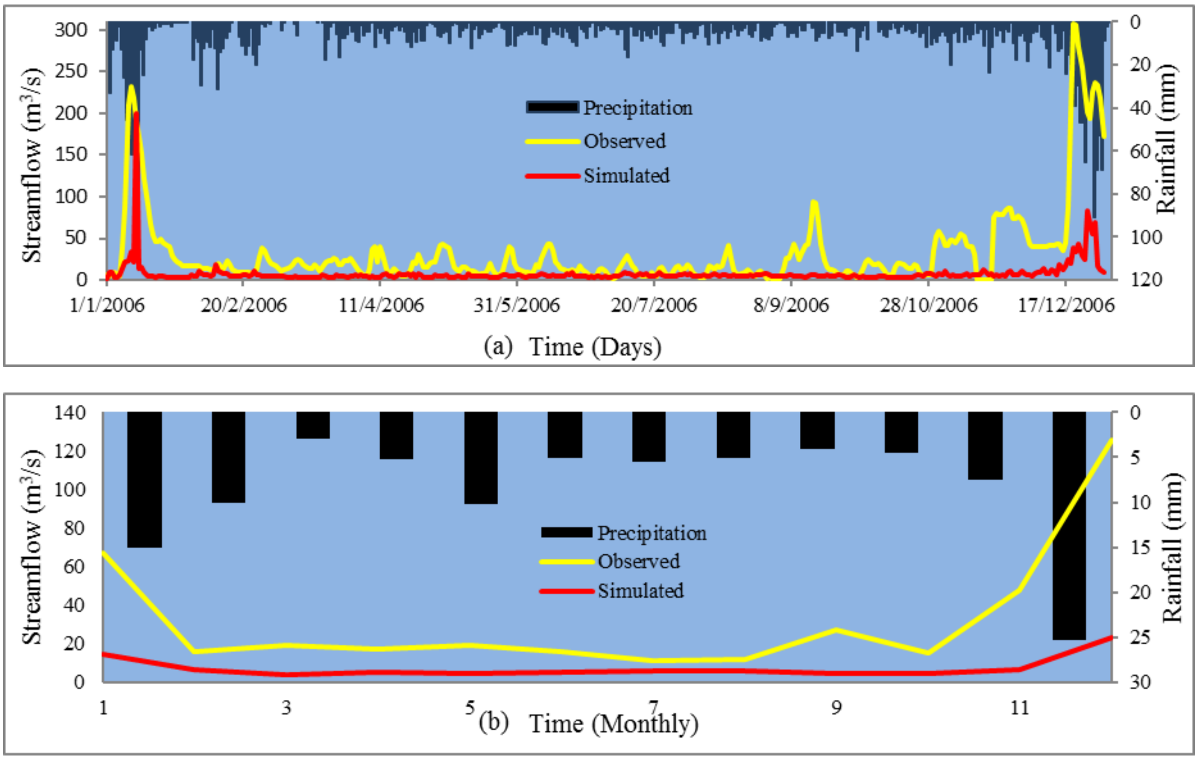


Fig. 3 Validation using observed streamflow at Rantau Panjang Station, (a) Daily (b) Monthly.

4. Result and Discussion

The performance of the public domain datasets for daily and monthly streamflow simulation by the SWAT model for the year 2006 are shown in Fig 3. Results for the daily and monthly streamflow simulation at Rantau Panjang station for the year 2006 indicate that NSE and RSR are in acceptable performance with values of 0.02, 1.01 and 0.21, 1.10, respectively. But, the value is classified as unsatisfactory acceptance (less than 0.5) [17]. The monthly scale performance is better compared to the daily timescale. The PB indicates an underestimation of bias of the modelling with a percentage of 77.07% for daily streamflow and 76.82% for monthly streamflow. There is a very good correlation looking at the standard regression statistical analysis where the degree of collinearity R2 for monthly streamflow in a value of 0.92, while only satisfactory for daily simulation with a value of 0.35.

Table 1 Statistical analysis result for daily and monthly simulated streamflow.

	NSE	R^2	PB(%)	RMSE (m^3s^{-1})	RSR	MAE(m^3s^{-1})
Daily	0.024	0.346	77.07	50	1.012	25.46
Monthly	0.205	0.923	76.82	37.04	1.098	25.23

Evaluating the results, the monthly streamflow simulation shows a better performance compared to the daily time scale, which means that public domain data sets are suitable to simulate the monthly streamflow for a tropical region. The main reason that result cannot reach acceptable performance classes with NSE higher than 0.65 [17] is that there is no sensitivity analysis and no calibration process done for the modelling. The main purpose of both processes is to maximize the performance of the model with a raise of the NSE value between simulated and observed streamflow [18]. Furthermore, precipitation is often cited as being the most important input parameter to any hydrological model [1], so inaccuracies of the precipitation datasets influence the success of the modelling [19]. NCEP space based precipitation datasets with coarse resolution is cited as the causes of unsatisfactory performance. Vu et al. [20] evaluate the performance of different space based precipitation products at a Vietnam River Basin and found that NCEP precipitation datasets did not show a good correlation for the daily streamflow modelling using SWAT. In a further study, evaluation of different types of remotely sensed precipitation products in SWAT model for Johor River Basin should be conducted to identify the best data for such conditions. Then such product could replace the NCEP precipitation data sets while maintaining other climate variables in order to improve the model performance.

5. Conclusion

This study presented the application of Internet data sets in streamflow modelling for a tropical region, Johor River Basin in the Malaysia using SWAT. The statistical analysis proves that these datasets provide an acceptable performance. However, the suitability is limited to monthly simulations rather than daily streamflow modelling. This study proved that public domain data sets are more suitable to simulate the monthly time scale in a tropical region. The results suggest that ground data sets that represent the actual earth surface are vital to accurate streamflow information extraction. This information is required particularly in flood forecasting and for dam construction projects. The sensitivity analysis and calibration procedure should be performed to ensure better quality and minimize the uncertainties of modelling. Besides that, the evaluation and selection of the best precipitation product increase the accuracy of the result.

The free availability of geo-referenced datasets to the public is the main concern of the Digital Earth [21]. The free Internet data sets for streamflow modelling can act as alternative approach when there is a lack of monitoring stations (ungauged basin) or budget problems arise. Besides that, this finding is important for trans-boundary river basin research where user may face difficulties in ground data set collection. The input of public domain data sets from Internet into hydrological model is a very important finding for educational purposes. This will help to promote, understand and get involved in the earth's environmental system or hydrologic cycle. It will support especially students in their research to obtain the necessary data in an easy, fast and digital manner anywhere in the world.

Acknowledgement

I express my special thanks to the Ministry of Higher Education in Malaysia for the Ph.D. scholarship and Universiti Teknologi Malaysia for providing the research program. The author sincerely thank to FAO-UNESCO, NASA, NCEP, and USGS-LCI for providing data. The author likes to acknowledge to the supervision and comments by Dr. Ab Latif, Dr. Christine Pohl and Zheng Duan.

References

[1] Kite G W and Pietroniro A 1996 *Hydrol. Sci. J* **41** 563-91
[2] Bourdin D R, Fleming S W and Stull R B 2012 *Atmos Ocean* **50** 507-36
[3] Ames D P, Horsburgh J S, Cao Y, kadlec J, Whiteaker T and Valentine D 2012 *Environ. Model. & Softw.* **37** 146-56
[4] Lacroix M, Kite G and Droogers P 2000 *Research Report 40 Colombo, Sri Lanka: Int. Water Mangement Institute*
[5] Arnold J R, Srinivasan R, Muttiah and Williams 1998 *J. Am. Water Resour. Assoc.* **34** 73-89
[6] Griensven A V, Akhtar M K, Haguma D, Sintayehu R, Schuol J, Abbaspour K C, Andel S J V and Price R K 2007 *Proc. 4th Int. SWAT Conf.* (Delft: The Netherlands)
[7] Kia M B, Pirasteh S, Pradhan B, Mahmud A R, Sulaiman W N A and Moradi A 2012 *Environ. Earth Sci.* **67** 251-64
[8] Global Land Cover 2000 database. European Commission, Joint Research Centre http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php
[9] Stibig H J, Beuchle R and Achard F 2003 *Int. J Remote Sens.* **24** 3651-62
[10] Food And Agriculture Organization of the United Nations (FAO-UNESCO) soil map v3.6 completed January 2003 http://www.fao.org/geonetwork/srv/en/metadata.show?id=14116
[11] Shrestha M S, Artan G A, Bajracharya S R and Sharma R 2008 *J Flood Risk Manag.* **1** 89-99

[12] Reuter H I, Nelson A and Jarvis A 2007 *Int. J. Geogr. Inf. Sci.* **21** 983-1008
[13] Jarvis A, Reuter H I, Nelson A and Guevara E 2008 *Hole-filled seamless SRTM data V4* International Centre for Tropical Agriculture (CIAT) http://srtm.csi.cgiar.org.
[14] Environmental Modeling Center National Centers for Environmental Prediction National Weather Service 2010. http://rda.ucar.edu/datasets/ds093.1
[15] Nash J E and Sutcliffe J V 1970 *J. Hydrol.* **10** 282-90
[16] Legates D R and McCabe G J 1999 *Water Resour. Res.* **35** 233-41
[17] Moriasi D N, Arnold J G, Liew M W, Binger R L, Harmel R D and Veith T 2007 *Trans. of the ASABE* **50** 885-900
[18] Bitew M M and Gebremichael M 2011 *Hydrol. and Earth System Sci.* **15** 1147-55
[19] Larson L W and Peck E L 1974 *Water Resour. Res.* **10** 857-63
[20] Vu M T, Raghavan S V and Liong S Y 2012 *Hydrol. and Earth System Sci.* **16** 2801-11
[21] Gore A. 1998 *Australian Surveyor* **43** 89-91