Geo-information Science and Remote Sensing

Thesis Report GIRS-2018-18

INTEGRATING TIME SERIES FOREST LOSS INTO STREAMFLOW PREDICTION BY RANDOM FOREST IN KEY WATERSHEDS OF THE PHILIPPINES

Arnan Araza

April 18, 2018

WAGENINGEN
UNIVERSITY & RESEARCH

INTEGRATING TIME SERIES FOREST LOSS INTO STREAMFLOW PREDICTION BY RANDOM FOREST IN KEY WATERSHEDS OF THE PHILIPPINES

Arnan Araza

Registration number 86 12 12 017 120

Supervisors:

Prof. Dr. Lars Hein
Prof. Dr. Martin Herold

A thesis submitted in partial fulfillment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

April 18, 2018
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2018-18
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Contents

1 Abstract

Water from watersheds or catch basins are being regulated by forest ecosystem and forest loss is affecting it negatively. Philippine government started accounting for "Ecosystem Services" (ES) where water regulation service is often quantified using data-intensive process-based models. The indicator of this ES is streamflow (l/sec) or river flows in watersheds often modeled using a single period land cover/land use input, with series of parameterization, resulting to a time series output.

This study performed an innovative Remote Sensing-dependent technique to predict streamflow at large scales using Random Forest (RF) regression within 6 major river basins in the Philippines. It integrated yearly forest loss pixels in predicting daily seasonal streamflow looking at evidences of forest loss effect in the context of water regulation service. A total of 58 physical and climatic covariates were assembled using 4 grouping-learning methods of "valuetable", or covariates pool of information, to assess how good RF is in predicting at varying degree of watershed information. The best model was used to predict at full-scale time series from 2000 to 2016 and applied to 6 validation sites without observed data.

The RF models learned better when combination of subwatershed valuetable is applied. The RF model also captured and learned from mixed information. Results revealed which subwatersheds have artificial water regulating structures like dams (regulated subwatersheds) which were outperformed by unregulated subwatersheds based on accuracy measures. The latter also showed rainfall-responsiveness or being linearly reactive to rainfall (i.e. high rainfall-high streamflow) in a daily basis. Streamflow predictions were not correlated with the number of training data and subwatershed size. The 6 validation sites, which are all unregulated, showed rainfall-responsiveness and good potential for upscaling.

Forest loss and its associated covariates (FLAC) were valuable explanatory based on permutation and removal measures. More importantly, the RF models were capable of learning forest loss effect to seasonal streamflow based on the following indicators: (1) decrease and increase in dry and wet predicted daily mean streamflow, respectively; (2) linear trends between forest loss and predicted seasonal streamflow according to regression and partial dependence plots; and (3) regulation of wet season streamflow by controlling outliers and reducing over-prediction. Moreover, an increase in forest loss rate resulted into decrease in streamflow during dry season and increase in peak flows during extreme rainy days for unregulated subwatersheds.

2 Introduction

Ecosystem service is defined as the direct and indirect contributions of ecosystems to human well-being (Sukhdev, 2008). Water from watersheds or catch basins are being used by half of Philippine agricultural lands and domestic water users (Philippine Statistics Authority, undated). There are 18 large watersheds or river basins and 142 critical watersheds in the country (River Basin Control Office, undated) that are at risk to human and climate change-induced threats. Almost three-fourths of the Philippine terrain are ch9aracterized by these watersheds. The remaining xha of forests regulate water from watersheds, one of the many intangible services provided by Philippine forests.

The water flows coming from interconnected river network of a watershed is termed as "streamflows" measured in water volume per unit time, usually at liters per second (l/sec). In hydrology, water availability from river systems can be quantified using streamflow as a measuring entity (Milly et al., 2005). Moreover, health of watersheds can be reflected by the quality and quantity of streamflows from both natural and human-induced conditions (Erwin and Hamilton, 2002) like land use changes (LUC). In the Philippines, one of the prominent LUC is caused by deforestation and it is affecting the water flows (Kummer, 1992). Studies exist how a deforested watershed is very different from a healthy watershed as indicated by streamflows (Aquino et al., 2014; Hurkmans et al., 2009; Ochoa et al., 2016; Rawlins et al., 2017).

Remote sensing (RS) is a powerful tool to quantify ecosystem services (Vargas et al., 2017). Remote sensing performs synoptic, spatially continuous, and regular observations over the globe (Maes et al., 2013). In a review paper, 5,920 studies were found using remote sensing applications for ES researches (de Araujo et al., 2015). A study by Feng et al. in 2010 concluded that RS-ES link happens in three ways: direct monitoring, indirect monitoring, and combined use with ecosystem models. For ES models in particular, RS-ES has essentially an "input-output" relationship where RS images serve as key inputs. For example, using land-use/land-cover (LULC) classifications from time-series of Landsat images, many studies were able to measure deforestation rates (Wang, 2006; Volante et al., 2012; Klepeis et al., 2013). Another example is a study by Shortridge et al. (2016) using agricultural land cover temporal changes to simulate streamflow.

In this decade, RS products with high spatial and temporal resolution are getting more accessible. For instance, the opening of the Landsat archive is used to analyze past and present global land and water change that paved way to products like the global forest loss (Hansen et al., 2013) and global surface water (Pekel et al., 2016). These kinds of accessible RS products further strengthens time series-based research and applications for large area monitoring of land and water dynamics (Broich et al., 2011). Numerous studies exist that capitalized time series RS data. An example is a near real-time monitoring of ecosystem disturbances methods which are being widely tested and standardized to address ecological issues (Verbesselt et al., 2012). Similarly, the study by Pasquarella et al. in 2016 assessed the potential of using Landsat time series to further the study of land cover characterization, vegetation phenology and landscape dynamics. Having easily accessible time series data can very advantageous to the RS-ES link by having historical up to near real-time RS inputs.

This study offers an innovative RS-dependent technique to predict streamflow in selected subwatersheds of Philippine major river basins. It integrates a time series forests loss input to a machine learning algorithm where its importance along with other predictors are assessed. Machine learning is perceived as an alternative to process-based models to address data and processing intensity issues

of the latter (Shortridge et al., 2016). Moreover, machine learning is considered as a non-parametric spatial modeling approach (Xu et al., 2016) that is free of series of parameterization unlike that of process-based modeling approach. Random forest, one of the best machine learning techniques, is applied successfully to various real-world classification and regression particularly on streamflow prediction by the following previous studies: Carlisle et al., 2011; Galelli & Castelletti, 2013; Rasouli et al. 2012; Reynolds & Shafroth, 2016; Sando & Chase, 2017; Shortridge et al., 2016; and Zhao et al., 2011. The study by Shortridge et al., (2016) found out that RF model had the lowest mean absolute error of streamflow prediction among other machine learning methods. By definition, the algorithm is a rule-based, non-parametric regression approach where randomized subsets are created (trees and branches), trained from separate bootstrapped data using small/random subset of predicting variables, and on a majority rule assign classes by decision tree votes (Breiman, 2001). A set of dynamic and static predictors or covariates are used for this study. The former consist primarily of weather data and forest loss with "associated covariates" or those parameters that are functions of forest loss. These are Curve Number and Manning's roughness coefficient which both characterize surface run-off or how slow or fast surface water moves to the rivers. The latter, on the other hand, include physical inputs like soil characteristics, elevation, and slope.

The study area shown in Figure 1 composes 6 major river basins in Luzon, Philippines namely Abra, Agno, Apayao-Abulug, Cagayan, Marikina, and Pampanga with an average size of (10,790km^2) per basin. More than half (60%) of the total basin area are forest lands by law according to River Basin Coordinating Council. Since 2000, the forest cover is reduced by 1,258km^2 (Hansen et al., 2013) at an alarming deforestation rate. The focus areas of the study include a total of 21 subwatersheds out of a data-driven selection seen also in figure below. The acronyms used and designated subwatershed names are shown in Annex A.

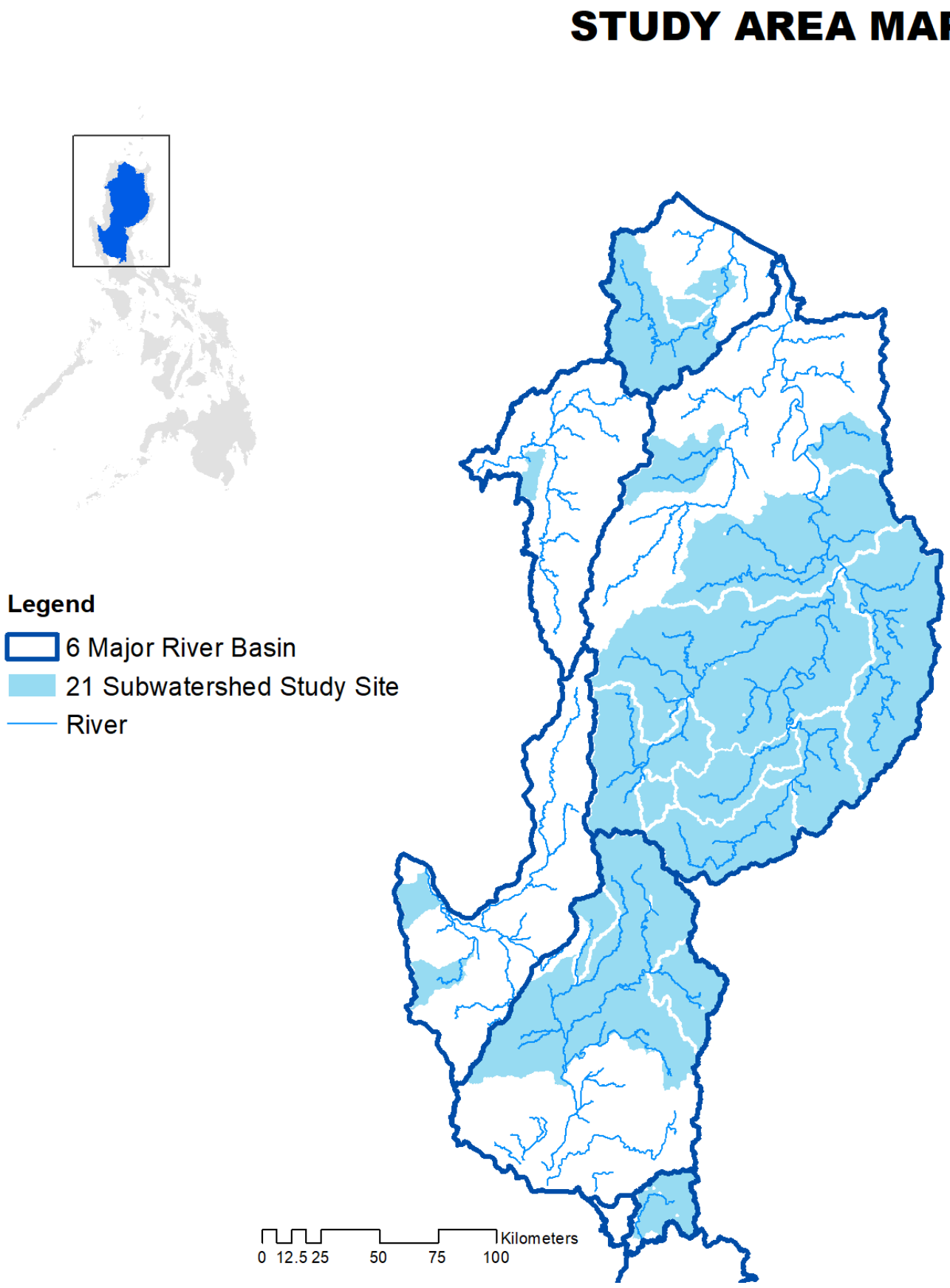


Figure 1: Study area overview.

3 Problem Statement

Philippine forest cover was almost half of its total land area during the 1960s (DENR-FMB, undated). However, logging companies abused it by rampant clear cutting from 1960 to 1990s. Even in the past 16 years, the country lost almost a million hectares of forests according to globalforestwatch.org. Losing forests linearly means decline in ecosystem services it provide (Pilot Ecosystem Account for Southern Palawan, 2016).

Quantifying ecosystem services could be complex depending on the type of service and its corresponding modeling approach/es. In theory, it contains non-linear dynamic processes, feedback mechanisms and control strategies to deal with complex ecosystem dynamics (Remme et al., 2014). Ecosystem services modeling is within the domain of social-ecological systems which is an integration of biophysical and social factors characterized by complexity (Levin et al., 2013). Moreover, ES quantification highly rely on scale, data availability, and technical expertise. In modeling water services, methodologies vary as there is no standard modeling blue-print to capture hydrological dynamics relative to ecosystem structure/landscape. There is a particular spatial-temporal dimensions requirement for hydrological modeling (Wang & Ren, 2008) which is dependent on the modeling objective and analysis scale.

Water services from watersheds is often quantified by modeling streamflows using process-based models. In other words, streamflows are simulated after series of parameter setup using a modeling tool or software. This technique often uses a single period LULC input in simulating time series outputs. In some cases, periodic LULC is implied in process-based models by using two LULC periods. This could be sufficient if the watershed is properly managed and protected (meaning no conversion is happening), but if not, fixing land use inputs into just one period may not reveal the true picture on streamflow simulations. Most apparent transformations in watersheds over the past 150 years is the change in land use and natural land cover (Mustard et al., 2012) meaning these changes should be accounted historically and regularly. Moreover, the continuity of the simulations from process-based programs might get altered as the process needs warm-up years to run smoothly (Arnold et al., 2012). There were some attempts to integrate yearly land use dynamics to process-based models but data availability and calibration efficiency became the limitations (Koch et al., 2012). In the Philippines, temporal resolution of land cover model inputs is a limitation as the official land cover is released every 5 years at least. Examples are studies by Principe (2012) and Alibuyog (2009) in the Philippines where a land cover of a certain year was used to simulate time series streamflow.

Theoretically, deforestation and degradation alters hydrologic system of watersheds (Ataroff et al., 2000). Forests can play a very important role in hydrology as it regulates water in dry months and act as sponge during extreme rain events like tropical storms (Calder, et al., 2008). Evidences exist how forest conversion can affect the lowland or service areas immediately like irrigation and domestic water stakeholders in dry months or a flood plain area during rainy season especially during extreme rainy days like typhoons (Aquino et al., 2014; Calder and Aylward, 2006; Rawlins et al., 2016; Sahin et al., 1996).

Despite the deforestation-related water issues, penalization are seldom and often missing. There is limited technical capability and political will that can enable policy implementation on the said matter. For example, the water-polluting mining area is closed for investigation, the deforested area is a private property, and the upland cultivations are too remote. This physical inaccessibility can be over-

come by remote sensing techniques as it can acquire land features without physical contact regularly. Being specific on when and where forest changes occur on a regular basis could be helpful legislatively.

Families in the upland are relying on traditional farming ever since. Most of them are still venturing to unsustainable farming system like slash and burn within their ancestral lands. The outcome of this land use practice appears in relatively small patches in a high resolution satellite image. Detecting this land use practice can be difficult if a coarser resolution like MODIS (250m) is to be used for land monitoring. Other forms of forest conversion due to small-scale mining and agriculture production (i.e. coconut) could also appear in patches and are better seen in a higher resolution RS image. Integrating these relatively small forest conversions in a hydrologic model can also contribute in analyzing forest loss effects to hydrology (Jones et al., 2009).

Most farmers in the Philippines are still practicing traditional rice farming where farm lands are either irrigated or rainfed with 2-3 and 1 cropping season, respectively. For the former, irrigation water are coming from streamflows (National Irrigation Authority, undated). Droughts are experienced in the past decade (2006 and 2012) within the study area which dried rivers and affected thousands of hectares of rice lands (Taguinod, 2010 and The Northern Forum, 2016). More than that, future climate and weather patterns are foreseen to be unpredictable due to climate change (PAGASA, 2016). Forecasting the streamflow that integrates the said extreme scenarios can help the government in planning and land management for the farmers.

To briefly address the abovementioned problems, this study puts high spatial and temporal deforestation input in the modeling process to predict present (and even future) streamflow. More significantly, the study highlights the link between forest loss and seasonal water flows which can aid in forest ecosystem accounting at large scales. These two rationale serve as its main added value among other studies with similar methodology but lack the said time series input and observation. Additionally, the methodology can offer an alternative to process-based models easing data-intensity requirement and flexibility to model watersheds at different sizes. Furthermore, it can increase pre-processing and modeling efficiency via processes automation with the use of ready-made packages, especially *randomForest*, in a programming interface. Lastly, study results can provide support to justify deforestation effects to water availability in forestry policies perspective.

4 Research objectives and research questions

4.1 Research objectives

The overall objective of the study is to integrate yearly forest loss in streamflow prediction and assess its implications to seasonal streamflow at varying watersheds.

Below are the three specific objectives of the thesis:
1. Create four (4) random forest (RF) models that include net forest loss time series (2000 to 2016) and assess the best model for time series streamflow prediction.
2. Assess the importance of forest loss and its associated covariates (FLAC) using three (3) accuracy indicators.
2. Assess the effects of forest loss to the predicted streamflow seasonally using hydrographs, historical flows, and regression results.

4.2 Research questions (RQ)

To answer RQ #1, covariates information were grouped in different ways to assess whether RF is capable of learning at varying degree of subwatershed information. Accuracy indicators in Section 4.2 were measured after and served as indicators.

To answer RQ #2, three measures were assessed with and without the forest loss and associated covariates (FLAC) to evaluate the effect of FLAC to the random forest models. These are permutation measure, removal measure, and partial dependence plots.

To answer RQ #3, evidences pertaining to the effect of forest loss to seasonal streamflow were investigated. Focus of analysis were subwatersheds with historical data and good performing models based on accuracy indicators.

The research questions of the study are the following:
1. Is RF able to predict streamflow well at varying degree of subwatershed information?
2. How do FLAC affect the RF model and predicted streamflow according to accuracy measures?
3. Can forest loss affect the predicted streamflow during wet and dry season?

5 Methodology

This thesis part narrates all the steps accomplished from data acquisition to results analysis. It has 4 main sections starting from data acquisition and pre-processing together with the methodology of the 3 research questions. The overall work flow diagram of the methodology is shown in Figure 2.

5.1 Data acquisition and pre-processing

Overview

This section basically pertains to the acquisition process with basic metadata of all data acquired and its preprocessing, often in steps, that preceded after. In general, the acquired data were all open source spatial raster images except for the weather and streamflow data. All spatial data underwent re-projection into UTM Zone 51N then mosaicking and masking raster operations since the study area is covered by several raster tiles. On the other hand, the weather and streamflow data were just formatted accordingly. At the end of all preprocessing, each data input was overlaid per subwatershed. Other intermediate preprocesses are explained per subsection in details. As a last note, most preprocessing steps were automated in R programming language as shown in Annex B.

DEM and DEM-derived data

Digital Elevation Model (DEM)-derived inputs namely elevation, slope, and watershed boundary were obtained from the Shuttle Radar Topography Mission Digital Elevation Model (SRTM-DEM) at 90m resolution from a personal copy originally acquired from https://earthexplorer.usgs.gov/. The choice of DEM was based on study area size which is relatively large therefore using the SRTM than a higher resolution could save processing time. Moreover, a handful of publications reported the use of the said DEM for hydro-modeling and online discussion users preferred it than its main counterparts because of less topographic errors (Ali, 2015). The DEM error known as "sinks" or the topographic holes of the DEM were filled first in a GIS interface (ArcMap) using *fill sinks* tool. That process was a common preliminary intervention to DEMs especially in hydrological modeling to get rid of erroneous water flow direction, accumulation, and watershed delineation (Van Remortel et al., 2001).

Elevation input was essentially the DEM itself in meters above sea level (MASL) units. The only processing to it was reclassification into 5 classes since its value range, being a continuous data type, was enormous to be a covariate each. The classes were derived by a *quantile* type of classification Using *reclassification* tool of ArcMap. This was chosen among 5 other reclassification tool to assure equal assignments per class, thus, not having empty classes or classes with too few or too many values.

The slope input in % rise was derived from the DEM using the *slope* tools in ArcMap. The product of the procedure was reclassified manually using the *reclassification* tool of the same software to have 5 slope classes that jive with the official slope classification of Philippines: 0-8%, 8-18%, 18-30%, 30-50%, and 50% above.

The watershed boundaries were derived from the DEM using a hydrologic extension tool in ArcMap.

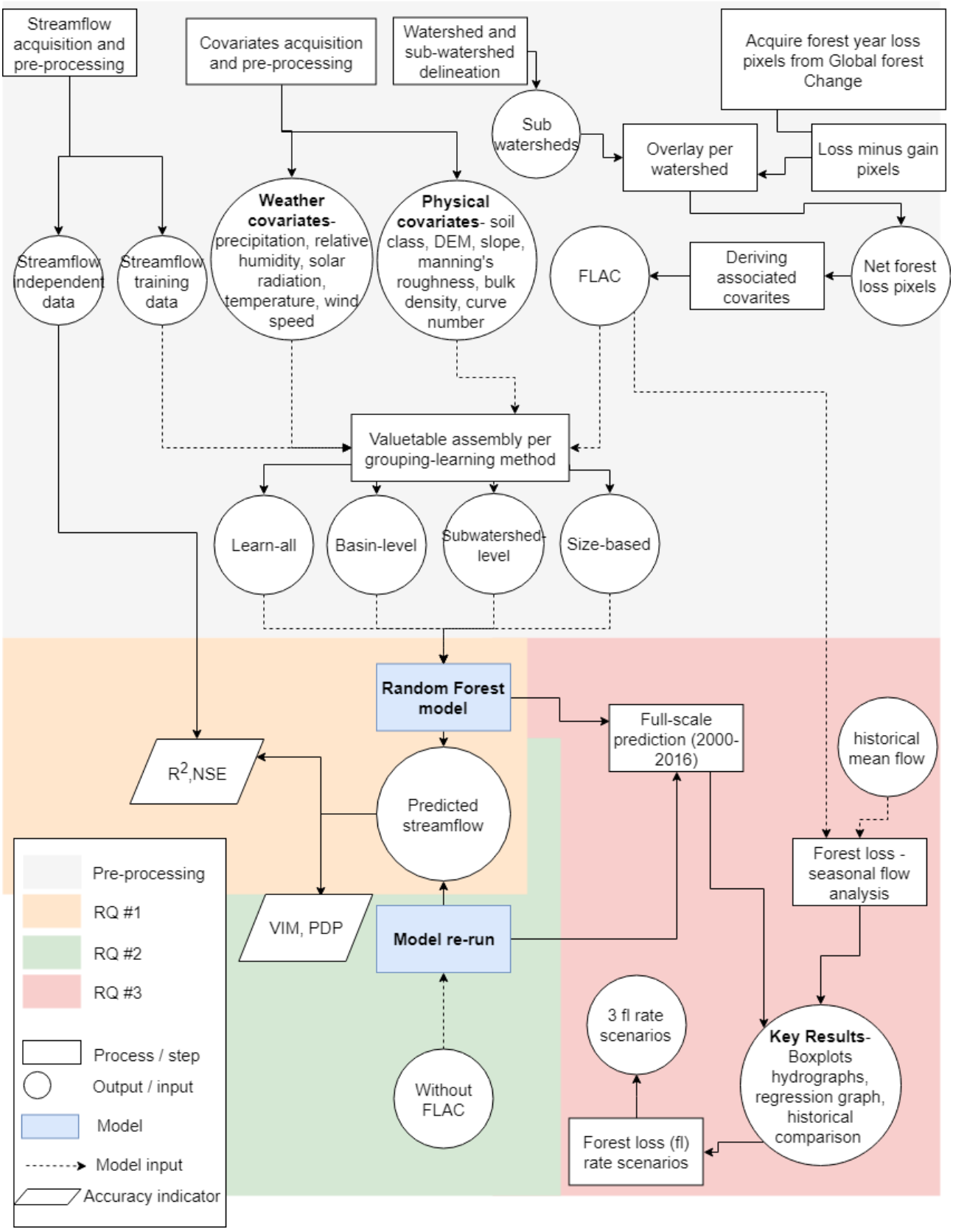


Figure 2: Schema showing the methodology of the thesis per methodology phase.

It was used because of a Graphical User Interface (GUI) availability defining the following necessary parameters to delineate watersheds: flow direction, flow accumulation, area, and inlet-outlet. The first three parameters were generated using default values and the outlet was manually selected for each discharge point. In the end, the 6 major river basins have 142 subwatersheds. The delineation was satisfactory since both small and large subwatersheds were among the 142 delineations. This is important since watershed size is one of the analysis angles of the study.

The elevation and slope inputs were overlaid with the subwatershed boundary by rasterizing the subwatershed polygon then being stacked together with the two rasters using *stack* function from *raster* package in R. Its values were derived using *getValues*, also from *raster* package, which gave the attributes of all raster layers in one data frame.

Forest loss and land cover

Yearly forest loss pixels were downloaded from Global Forest Change at https://earthenginepartners.appspot.com (Hansen et al., 2013) which was Landsat-derived at 30m resolution. The study area was covered by two raster tiles so mosaicking was needed and was done in R. Forest gain pixels which happened not to be yearly were subtracted to the forest loss pixels to derive the net change pixels. In other words, those deforested earlier which happened to regenerate after the end period in 2016 was not included as a forest loss pixel. This step was accomplished by masking the 2016 gain pixels to the 16-year forest loss pixels. The purpose of this was to not amplify forest losses and also account for the forest gain impact to hydrology. Forest gain in the study area, however, was perceived to be relatively small compared to what were lost based on periodic land cover comparisons.

On the other hand, the baseline land cover data was obtained from global land cover (glc30) also at 30m resolution from the site http://www.globallandcover.com. The study area was covered by 2 raster tiles which also underwent mosaicking and masking procedures, respectively. This land cover dataset was chosen to be consistent with the spatial resolution of 30m, same with the forest loss pixels and also both Landsat-sourced.

There was an attempt to overlay forest loss and land cover inputs together with the subwatersheds in R, however, the computer power demand was too much. As such, the operation shifted to ArcGIS for mosaicking and masking per individual input then overlaying the two with subwatersheds. Reclassification was found out to improve efficiency by quickly getting rid of undesired pixels of the land cover image rather than directly doing the clipping process.

Soil data

High resolution soil classification from World Soil Information (ISRIC) via soilgrids.org and its key parameters bulk density, texture (sand, silt, clay), and soil depth were downloaded. These parameters affect water infiltration and regulation, the two important soil hydrologic parameter (Calder, 2008). For so long, coarse soil inputs (1km or more) were used for hydro-modeling but the highly variable ISRIC soil grids at 250m can make soil inputs more relevant.

Preprocessing the soil inputs required mosaicking of tiles since it took four tiles to cover the whole study area. This step was done in R using *mosaic* function from *raster* package. The mosaicked

soil raster was masked after using the polygon of the whole study area. The raster needed to be re-projected first using *projectRaster* function then needed to be of the same extent with the subwatershed using *crop* function before doing the actual masking using the *mask* function. A function was made to automate this process for each soil input using the soil code as input variable. The mosaicked soil rasters were reclassified in a GIS interface using *quantile* type of classification resulting into 5 classes for each soil input. This was done to avoid crowded class covariates and save computing time since a soil input had 60 classes originally. Soil data were combined with the subwatershed raster then values were extracted using same procedure as the raster inputs above.

Streamflow data

The heart and soul among the data inputs were the actual streamflow data from selected river stations of the Philippines. All records within year 2000-2016 from any available stations within the study area of the Department of Public Works and Highways (DPWH) - Bureau of Designs were secured for model training/fitting. The said office records daily river flows from river monitoring stations in the Philippines for infrastructure development. Every local river station is being tasked to store daily streamflows then the main office are integrating it. Historical data from 1980s to 1999 was also acquired to derive historical mean water flows needed for detecting streamflow anomalies in Research Question 3, however, the data was only for 4 stations.

Most of the streamflow data were below 2010 because the bureau is currently updating their database and most of the current data are still in hard copies subjected for validation. However, the acquired data has an average of 2,461 observations per subwatershed and is seemed to be sufficient for the thesis' purpose. It should be carefully noted that the choice of subwatersheds were dependent on the available streamflow data. After the acquisition, 21 river stations fell into the study area therefore having 21 subwatersheds to be officially called "subwatersheds-of-interests".

The data were formatted accordingly taking-off from the bureau's formal template. Each river station was re-formatted into individual comma separated value (csv) sheets with heading of the watershed and river station name. This was purposely done for proper subwatershed-observed data pairing that happened afterwards during the assembly of the covariates table. The streamflow data had minimal NAs or Not Available but since it serve as the "observed data" for model training, no imputation/gap-filling was done as NAs data were just simply omitted.

Weather data

A stretch of 16-year time series (2000-2016) weather data was acquired from 12 stations of Philippine Atmospheric Geophysical and Astronomical Service Administration (PAGASA), the official weather institution of the Philippines. These are precipitation, relative humidity, wind speed, minimum temperature, and maximum temperature. The weather data were recorded manually from local stations then were collated at PAGASA main office. Manual encoding with high quality checking were done by the office data administrators to put the data in the weather database. Though global weather data were extensively available in the study area, local weather data were favored for a more localized and higher resolution data.

Since weather stations in the Philippines contained gaps from station malfunctioning, preprocessing

the missing value using imputation packages in R was needed. However, some stations were not operational for months. As such, data gaps were predicted using Inverse Distance Weights (IDW) spatial interpolation method among the stations under the same climate type. Prediction using regression from adjacent weather stations was the first attempt to fill the gaps but low correlation was observed due to high climatic variability. Another attempt was to do an ordinary kriging, however, the data points were insufficient to account for short distance variability and to create a well fitted variogram. Moreover, the climate variability might need a block kriging approach to really account for the climate types, such work can be a stand-alone study. The gap-free weather tables were binded in one data frame with station names as the column heading not just for proper formatting but also for some filtering technique later on during all covariates assembly.

Other data

The recent versions of the master plans for the 6 major river basins within the study area were downloaded from the River Basin Coordinating Office of the Philippines or RBCO. These documents were used as supplemental information for results analysis. Moreover, some intermediate results like forest loss rates and conversion trends, were validated through these plans.

The official land cover data of the Philippines for two periods (2003 and 2010) were acquired also from a personal copy which were originally generated by the official mapping authority of the Philippines. According to its metadata, the land covers at 30m and 10m respectively, had significant classification accuracy outcome. The land covers served as validation inputs for the forest conversion test case which will be explained in Research Question #3 methodology section.

5.2 Streamflow prediction for Research Question #1

Overview

The *randomForest* and *ranger* R packages were used to produce random forest models to predict streamflow as it was used in previous studies (Carlisle et al., 2011; Galelli & Castelletti, 2013; Shortridge et al., 2016; Zhao et al., 2011). As an overview, every covariate was assembled into a *valuetable* or simply the collection of all covariates values/information. A subset of it containing of 90% the *valuetable* became the "training data" for model fitting while the other 10% was used for model testing. The RF models were used to predict streamflow using *predict* function in R and its predictions were assessed versus the observed data using accuracy measures. All scripts used at this point are too long to be placed in the annex section. However, all scripts can be accessed at *insert hyperlink* depository page.

5.2.1 Covariates

There were a total of 58 covariates coming from the 18 inputs in Table 3 which were categorized as static and dynamic covariates to give emphasis in temporal resolution. In concept, the static covariates taught the model on the watershed's geophysical features while the dynamic variables were more temporally important. The dynamic covariates included weather data and FLACs while the rest were considered as static. Layout maps of the selected main inputs are shown in Figure 3

The covariates could be as many as 200 given the open source data available. However, having multiple predictors that are highly correlated can be an issue in streamflow prediction (Kroll et al., 2004). As such, other soil and vegetation static parameters were left out. The chosen soil and land cover covariates depicted hydro and land dynamics specifically water infiltration and surface run-off more significantly (Calder, 2008).

Forest loss and associated covariates (FLAC)

The forest loss was one of the key predictors as it affected two land cover covariates and 2 other covariates aside from being a covariate itself.

The net forest loss integration to each subwatershed was automated taking off from the pre-processed forest loss data with land cover and subwatershed information. It essentially provided hectares of forest loss in every subwatershed that enabled a cumulative decline of baseline forest cover per year. First, yearly forest loss was replicated on a daily basis by creating sequence of days from 2000 to 2016 using functions *seq* and *as.Date* in R then using the function *rep*. Three sub functions were created to make this happen. A function (1) that generated its cumulative difference per year was made then supplied to a function (2) that deducted it yearly from the baseline land cover per subwatershed was run yielding a 16-year output per subwatershed. Lastly, a replacement function (3) into the *valuetable* was created to integrate the second function's output. Main R functions used were *lapply* and *mapply* which enabled to do functions over a series of years.

To train the model on land conversion effects, a separate covariate for land cover gain were made for 4 subwatersheds. In other words, this covariate is exactly the opposite of forest loss that changes yearly alongside the associated covariates. These were pre-selected subwatersheds based on the significance of forest loss indicated by forest loss rate in Table 1. Moreover, these areas depicted different type of conversion from forest to grass, forest to barren, and forest to croplands. Another reason for selecting the areas was validation materials availability as what are reported to the basin master plans were considered. More details of this are shown in the next main section. To integrate this to the *valuetable*, the conversions were tabulated and put to the first integration function that deducted the forest loss to forest cover and adds it to the known conversion.

Table 1: Rate of forest loss per subwatershed per year (source:Hansen et al., 2013)

Basin	Forest loss rate	Basin	Forest loss rate
aarb_a	12.5%	crb_p	7.6%
aarb_n	4.0%	crb_s	5.5%
abrb_s	0.8%	crb_t	5.0%
arb_b	3.6%	crb_u	8.2%
arb_c	6.5%	mrb_s	5.6%
crb_a	11.7%	prb_b	3.9%
crb_be	3.4%	prb_c	4.2%
crb_bu	8.1%	prb_a	4.2%
crb_d	6.9%	prb_p	3.7%
crb_j	7.9%	prb_r	2.3%
crb_m	4.2%		

Reactive also to forest loss were the associated covariates Curve Number (CN) and Manning's roughness coefficient as both are functions of forest loss. The parameters are essentially surface runoff

coefficients per land cover which change whenever forest conversion happens. Moreover, these are hydrologic sensitive parameters (Grimaldi et al., 2013) which can necessitate the model to learn land conversion. Each land cover were assigned (Table 2) to its standard CN and Manning's coefficient value from United States Department of Agriculture (USDA), 1989. The same procedure of cumulative difference and integration to the main *valuetable* was applied for both covariates.

Table 2: Curve Number and Manning's coefficient values used per land cover (source: USDA).

Land cover	Curve Number	Manning's coefficient
Forest	60	0.6
Perennial crop	79	0.22
Annual crop	81	0.17
Grassland	80	0.15
Wooded grassland	76	0.3
Barren land	91	0.01
Built-up	90	0.01
Water	70	0
Wetland	70	0

Weather covariates

Each of the following daily weather data were standalone covariates: precipitation, minimum temperature, maximum temperature, relative humidity, and wind speed. But aside from these, four (4) more weather covariates were derived to strengthen the highly variable weather pattern of the study areas. The mean precipitation per month, mean precipitation per week, maximum precipitation per week, and maximum precipitation in a day were added as separate covariates similar to the study by Reynolds and Shafroth in 2016 for a random forest-based streamflow prediction in a highly seasonal watershed. The first three were essentially a proxy to soil saturation effect from series of rainy days since knowing the maximum rainfall in a week could help the model distinguish random rainy days to rainy season. The maximum rain of the day was added to supplement model learning on peak flows of extreme rainy days especially typhoons and monsoons. However, due to data limitation, maximum rain per day covariate was tested for 1 subwatershed only.

Static covariates

Soil covariates, slope, and elevation were assembled in a script. Essentially, these were supplemental covariates to train the model how the physical environment affect the water flows especially when it comes to water infiltration contributed by soil texture, bulk density and soil depth covariates. The elevation helped by sufficing the model that highly elevated areas are also high-rainfall areas. The slope affected the model by contributing to the lag time effect or the time difference of the peak rainfall and peak flow. Combining these covariates resulted into 35 predictors which is 65% of the all covariates.

Other covariate

A "month" covariate was added due to climate and weather seasonality of the study areas. This way the model learned which months were dry and rainy in a 16-year stretch. Another covariate that distinguished each subwatershed was added (subwatershed code) that served as "labels" per subwatershed during *valuetable* grouping.

Table 3: All streamflow covariates/predictors for the random forest models.

No.	Predictor	Class/es	Unit	Nature	Time nature
1	Daily Rainfall	1	millimeter	climatic	dynamic
2	Mean Daily Temperature	1	°C	climatic	dynamic
3	Wind Speed	1	m/s	climatic	dynamic
4	Relative Humidity	1	%	climatic	dynamic
5	Mean Monthly Rainfall	1	%	climatic	dynamic
6	Max Monthly Rainfall	1	%	climatic	dynamic
7	Max Weekly Rainfall	1	%	climatic	dynamic
8	Soil types	5	pixel count	physical	static
9	Bulk Density	5	kilogram	physical	static
10	Soil depth	5	cm	physical	static
11	Curve Number	1	coefficient	hydrologic	dynamic
12	Manning's Roughness	1	coefficient	hydrologic	dynamic
13	Forest loss	1	pixel count	physical	dynamic
14	Land cover gain	1	pixel count	physical	dynamic
15	Other land covers	6	pixel count	physical	static
16	DEM	5	meters	physical	static
17	Slope	5	%rise	physical	static
18	Month	1	1-12	categorical	static
19	Subwatershed code	1	1-21	categorical	static

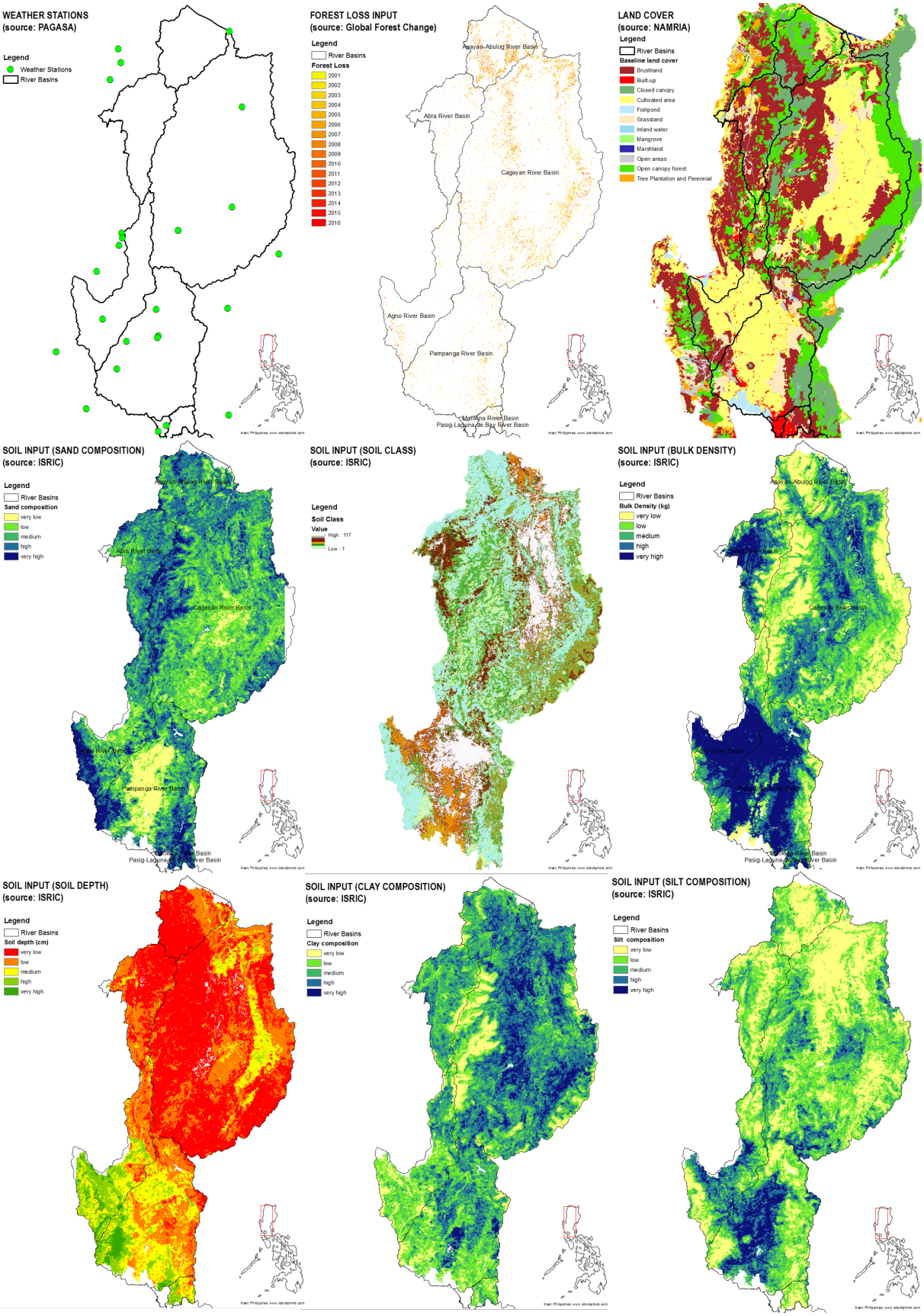


Figure 3: Key spatial inputs used as covariates.

5.2.2 Valuetables

All covariates were assembled into one main table/data frame termed as *valuetable* per subwatershed. The *valuetables* underwent a specific grouping-learning method before being supplied to the random forest model. These two-way process were done using a long R script with a combination of spatial and data frame manipulation packages namely r*aster, rgdal, dplyr, plyr, data.table, reshape2, varhandle, xts, Hmisc, splitstackshape*.

Assembly

The main function to assemble the *valuetable* took subwatershed name and a conversion boolean string as arguments. Series of scripting steps also shown in Annex B were accomplished to assemble a *valuetable*. An if-else statement was included also if the subwatershed had data on conversion since that variable was applicable only to 4 subwatersheds as explained above.

First, all individual covariate tables were loaded, all had subwatershed information except for the weather and streamflow. To deal with that, subset-pairing scripts were written. A subset method was done using *grep* function which selected columns based on respective headings of subwatersheds. To determine which is which for the subwatersheds, the *filter* function was used to predetermined subwatershed numbers which were assigned during the watershed delineation. The basis of determining the numbers were based on the discharge point location. Querying of the 21 out of 142 subwatersheds was done ArcMap's model builder because of its querying advantages and also for ease of map layout afterwards. Moreover, the basis for selecting the weather stations was also based on weather station location. The nearest stations within the subwatershed was selected. This is also the selection method of process-based models like SWAT (SWAT Documentation, 2012). In some cases, a thiessen polygon-based weather data assignment can be derived especially for smaller watersheds with numerous weather stations and for more in-depth goal like for flood forecasting. In this study, a 1:1 weather station-subwatershed was perceived to be sufficient given the relatively huge study area and amount of weather data acquired.

A table merging scheme was accomplished next after getting the values for the desired subwatershed. Control measures, including removal of NAs, to assure that *valuetable* had a total of 6210 days (2000-2016) was made. At that point, proper heading name was accomplished. The merging of individual *valuetables* was done lastly using the SW codes assigned from the watershed delineation.

Grouping-learning methods

The training data were assembled and trained in 4 different ways to assess how RF models are affected by watershed variability. Moreover, the grouping-learning method also served as the basis to choose which method is the ideal to use for the full-scale streamflow prediction from 2000 to 2016. The 4 kinds of *valuetable* grouping were: (1) "learn-all", (2) basin-level, (3) subwatershed-level, and (4) size-based. The first grouping was built by combining the 21 *valuetables* into one collated table resulting into 51,690 observations run in one model. The second method or the basin-level method assigned each of the 21 subwatersheds on its corresponding 6 major river basins resulting into 6 models. For the third method, each subwatershed was trained individually resulting into 21 models. Lastly, the *valuetables* were arranged per SW size resulting into 2 sub-models. Those watersheds below and above 100,000 ha were assigned to small and large model, respectively. Shown in Figure 4 is a diagram on how subwatersheds were arranged.

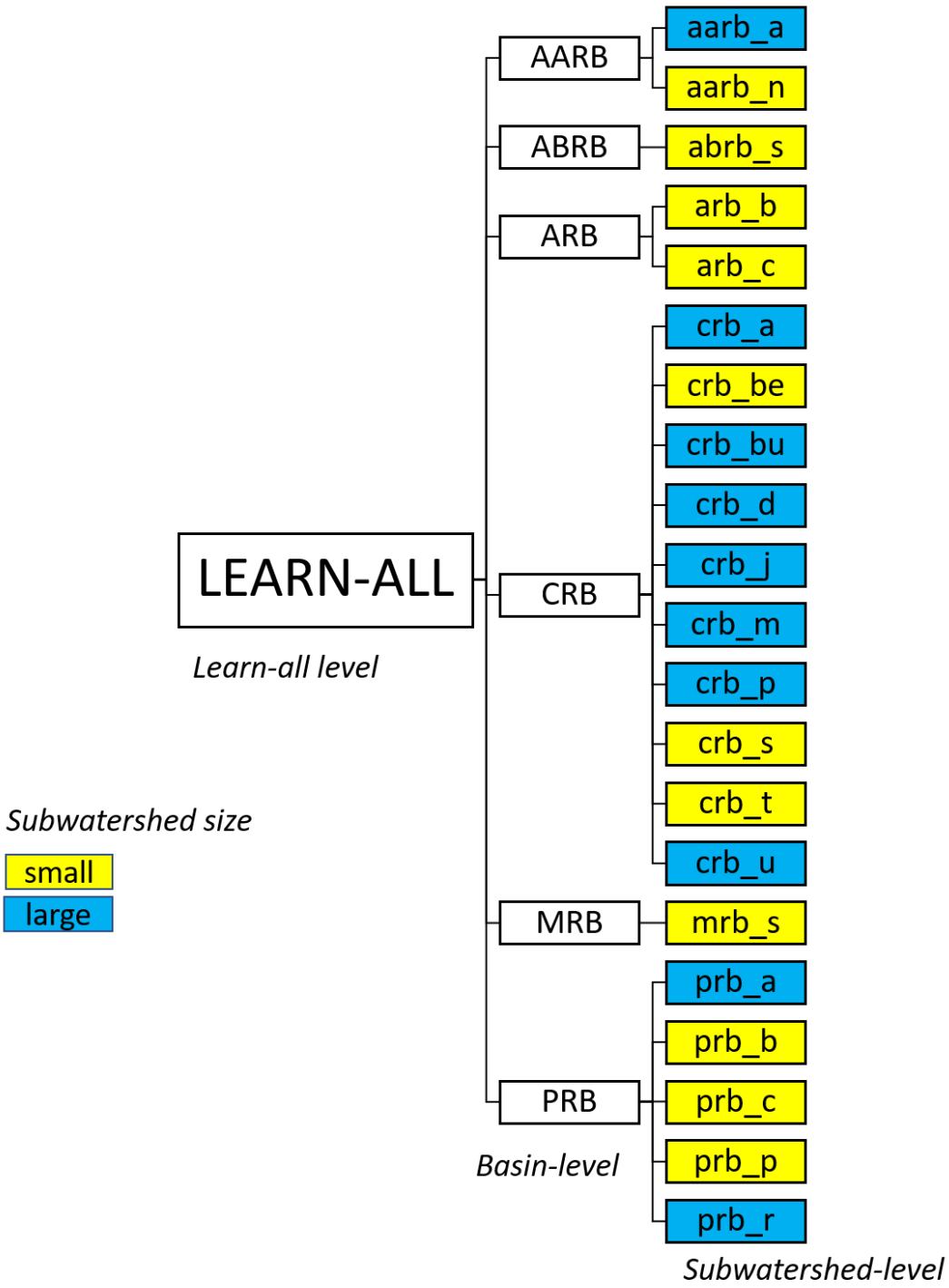


Figure 4: Hierarchy of *valuetable* grouping-learning methods used for random forest training.

A script to automate the grouping-learning methods was made not only to group the *valuetables* but also to facilitate the data division into training and testing datasets afterwards. It took grouping name, training data percentage, and conversion hint as script variables.

5.2.3 Random forest model

The algorithm

Random forest algorithm is an ensemble type of bagging or averaging method that reduces the variance of individual trees by randomly selecting many trees from the dataset, and averaging them. In bagging, successive trees do not depend on earlier trees where each is independently constructed using a bootstrap sample of the data set (Liaw and Wiener, 2002).

The RF algorithm proposed by Breiman (2001) adds a layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed.In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other machine learning algorithms and is robust against overfitting (Breiman, 2001). Each tree undergoes a simple binary decision and the final decision of each tree is averaged among the rest to get the final predicted value. A simple diagram of RF algorithm is shown in Figure 3.

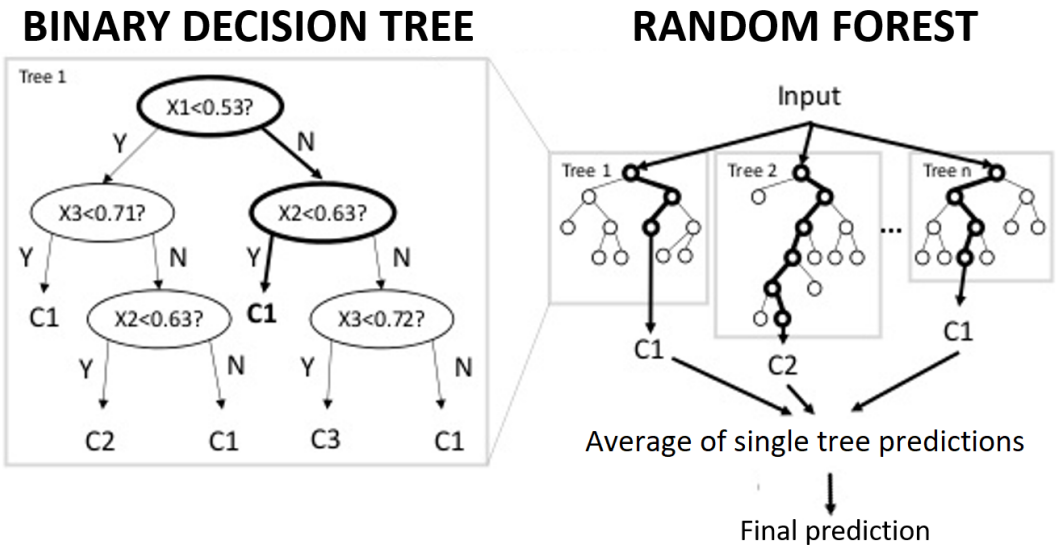


Figure 5: Diagram on how random forest algorithm is implemented (source: Nakahara, 2017).

Parameters

Two optimization methods were tested for key RF variables *mtry* and number of trees (*ntrees*). The former is the number of variables randomly sampled as candidates at each split while the latter is the number of subsets/trees to be grown. The optimization was to assure robust sub-setting leading to lower Out of Bag (OOB) error given a manageable computation time. The OOB is essentially a held-out data from the training data itself internal to the RF algorithm to produce initial validation measure of RF model. Moreover, *mtry* and *ntrees* parameters were perceived to be the most important for regression according to Breiman, 2001. Increasing the *mtry* is favorable but the trade-off is computation time (Breiman, 2001). Moreover, there is a "leveling-off" or a threshold where the OOB error will stabilize and not change despite increasing the parameter values (Sumatiphala, 2014). As such, the 3rd mtry value was chosen. The second optimization of key RF variables was done using *mlr* package *tuneParams* function. A range of +-10 to the 3rd *mtry* value given by function *tuneRF* were tested. Though slower than the first optimization method by far, this was done to see more details on how optimization was done and also to counter-check the first optimization method. This

became an intermediate experiment also as no existing study was made to compare the optimization methods but just user insights in related online forums. However, the only downside of optimizing was the computation time as it took longer hours to accomplish.

After tuning, the following were the optimized parameters with OOB error rate: mtry = 25, number of trees = 2500, nodes = 5.

Training data %

The percentage of training data were experimented for the basin-level grouping to determine which training-testing ratio was best to use for whatever random forest model. Moreover, it also revealed which basin was more sensitive to training data. As it appeared, the training % that attained the highest accuracy was 90% for all basins Table 4.

Table 4: Accuracy results per % training data used.

Basin	No. of observation	R^2 per training data %		
		90%	80%	70%
AARB	6954	0.64	0.62	0.60
ABRB	1461	0.44	0.40	0.45
ARB	5113	0.80	0.42	0.30
CRB	27402	0.61	0.54	0.52
MRB	2191	0.65	0.49	0.42
PRB	8570	0.67	0.62	0.66

The relationship among number of observation and watershed size with R^2 was also assessed using to investigate if it influences model results. Climatic and some biophysical information per site were also supplemented to support analysis.

Artificial high correlation which is a common sign of model over-fitting was avoided by controlling the training data attained using the sample function in R. This was done by activating *set.seed* random generator first before dividing the dataset into training and testing. The function is an integer vector, containing the random number generator state for random number generation in R (R documentation, undated). This was done since sampling is repeated twice in two processes during training and predicting and there were tendencies that the sample function chose different training observation when run twice.

Models

The grouped *valuetables* were run using *ranger* function for faster implementation of random forest in R developed and documented by Wright and Ziegler in 2015 except for subwatershed-level grouping which used *randomForest*. A comparison study was made by the said developers comparing ranger with other random forest ensemble methods. It was found out that ranger is the fastest and most computational-memory efficient implementation of random forest. It automatically assigns the processing to computer cores while taking the same computer memory. This was proven by an intermediate test at the basin level whereas virtually the same accuracy results were attained by the *ranger* and *randomForest* models. However, it did not work when implemented inside a function for sequential subwatershed processing, so *randomForest* function was used per subwatershed which took

way longer to run.

The rule of thumb was to have one RF model per *valuetable* grouping method. The product of the *valuetable* grouping script was fed to the RF models that were setup using the optimized parameters. Each model training was explicit in every grouping method. Running the RF models was straightforward but a separate script was still formulated to automate the sequence of variable optimization and model training. Moreover, the script automatically plotted the following results: optimization graph, variable importance measures (VIM), partial dependence plots (PDP). These intermediate results will be discussed more in the next section.

5.2.4 Accuracy assessment

Accuracy of the models was assessed using the predicted streamflow from the held-out data. Two relative measures were used for easier accuracy unit comparability also. A simple linear regression using *lm* package in R was used to derive the widely-used pearson's coefficient R^2. A threshold of above 0.60 R^2 was used to declare significant results. The second accuracy measure used NashâĂŞSutcliffe Efficiency (NSE) to assess the predictive power of hydrological models especially when used for prediction. It was computed using the *nse* function of *hydroGOF* package in R. Literature differ from judging the NSE performance into good or bad. Researchers at USGS reported 0.5 and above NSE represents a good fit for streamflow conditions (Christiansen 2012). Similarly, the SWAT calibration and validation documentation affirmed that threshold as good. A list of published NSE performance judgment was reported in a study by Moriasi et al. in 2005 (Table 5).

The Root Mean Square Error (RMSE) was not included as an accuracy indicator because RMSE is an absolute measure. The subwatersheds differ a lot in streamflow values because of varying sizes and characteristics, as such, absolute error comparisons is not advisable.

5.2.5 Comparison with other models

Secondary open source results of hydro modeling from the basins used for this study were compared with the thesis results. Most of the information came from the master plans of the river basins. It should be noted, however, that results of these secondary information were also modeling output with different methodology, temporal scale, and resulting units. As such, a simple regression to capture results trend was made between the thesis results and the secondary results. Capturing "similarity"

Table 5: NSE satisfactory values from various sources (source: Moriasi et al., 2005)

Model	NSE Value	Performance Rating
HSPF	>0.80	Satisfactory
APEX	>0.40	Satisfactory
SAC-SMA	<0.70	Poor
SAC-SMA	>0.80	Efficient
DHM	>0.75	Good
DHM	0.36 to 0.75	Satisfactory
DHM	<0.36	Unsatisfactory
SWAT	0.54 to 0.65	Adequate
SWAT	>0.50	Satisfactory
SWAT and HSPF	>0.65	Satisfactory

by regressing two modeling outputs can be sufficient just for comparison purposes.

Three sets of monthly average streamflow for basins aarb (master plan) and mrb (master plan and other study) were obtained (Table 6). Two were process-based model results while the other one was based on regression. The temporal scale of aarb included a historical scope which could have implications when compared to the results of this study.

Table 6: Other hydrologic model results within the study area of this thesis.

Basin	Source	Modeling Technique	Software	Temporal scale
aarb	master plan	process-based	SWAT	1980-2012
mrb	master plan	regression	unknown	2005-2010
mrb	other study	process-based	SWAT	2005-2010

5.2.6 Validation

A separate validation set shown in Figure 6 was run at full-scale (2000-2016) in order to evaluate if the models perform well in outside subwatersheds or those with no observed data but still within the 6 major river basin study area. There were 6 subwatersheds selected based on distribution and representation of the 6 basins. The naming of the validation sites correspond to it basin (first 3 or 4 letters) and size (last letter).

The predicted streamflow from the validation areas were judged based on seasonality since observed data were missing to perform accuracy measures.

5.3 Forest loss and associated covariates effect to the models for Research Question #2

5.3.1 Permutation and removal measures

The models with different *valuetable* assembly-training methods were tested on how FLAC affected the models. The FLAC were assessed using two covariate manipulating measures: permutation and removal. For the former, each modelâĂŹs importance measure based on post-permutation accuracy was compared. That measure gives the average accuracy difference on pre and post permutation between the RF trees. The higher the value, the more it influences the model by decreasing unexplained variance or simply increasing accuracy. Value ranges vary depending on the number of training samples so basins that have similar ranges were grouped for plotting purposes. Moreover, for the 21 subwatersheds, an extra graph containing %MSE change when covariate is permuted were included and assessed. This was at the subwatershed-level method because remember that it used the *randomForest* function instead of *ranger*. For the latter, each FLAC was removed and all models were re-run to check how the accuracy indicators (R^2, NSE, and MSE) changed after. The regression results were examined carefully to check for prediction changes and outliers. Observed and predicted line graphs were generated to visually aid the comparison.

The ranking of the covariates based on the importance measure was graphed to check how the FLAC performed with other covariates. This was done for all the models at the basin-level. The %breakdown was shown in both table and graph of at least 13 out of 58 covariates.

Lastly, the discrepancy from the observed data was compared for simulated flows with and without FLAC by computing for the % change using the simple equation below. This step was done to make a relative comparison, in line graphs, among basins using monthly streamflow averages and to check for over and under prediction of the predicted flows.

$$\%change = \frac{\text{simulated - observed}}{\text{observed}} \times 100$$

5.3.2 Partial Dependence Plots

Partial dependence plots (PDP) which measures how a particular covariate influence the prediction of the dependent variable by changing the value of one covariate while leaving all other covariates constant also served as an accuracy indicator. The PDPs were in the form of line graphs which are essentially similar to regression graphs that depict linearity among the two variables. The PDPs were derived using the *generatePartialDependenceData* function of *mlr* package in R for the full-learning and basin models. This is because there is still no exclusive PDP generator for *ranger* function. But for the subwatersheds that used *randomForest*, function *partialPlot* was used which is more straightforward. The PDPs of FLACs were assessed by comparing graphs and tables among the 21 subwatersheds, getting the common trend among the subwatersheds, and explaining the trend generated by each FLAC. The PDP showed the covariate of interest in the x-axis while streamflow in the

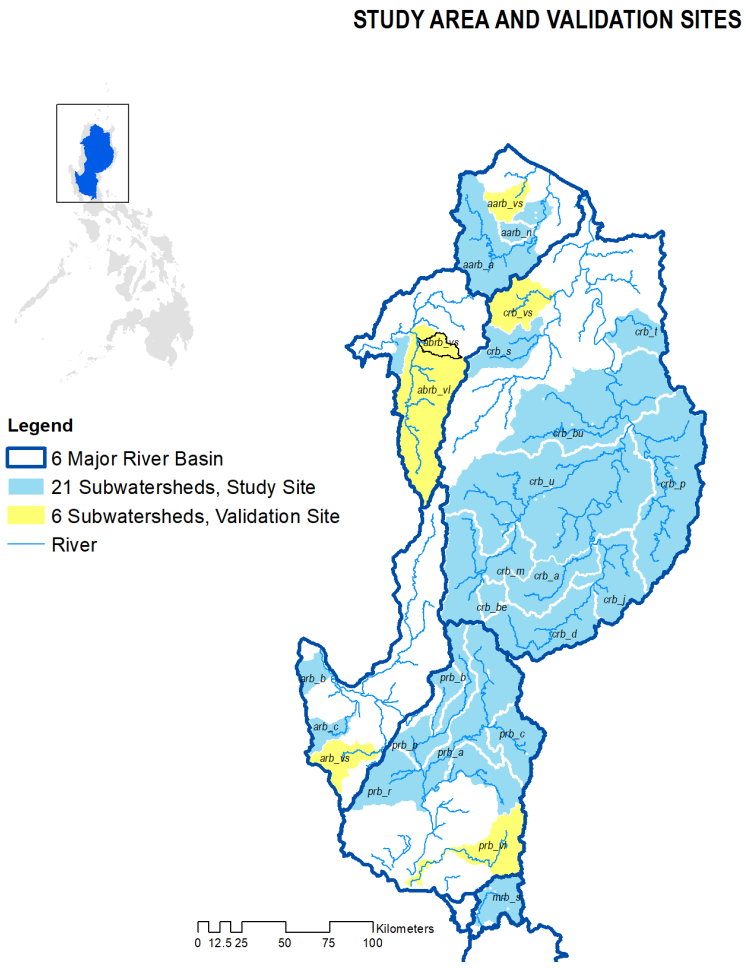


Figure 6: Map showing study area and validation sites.

y-axis. The trends of the PDP line graph gave hints on how the specific covariate values affected specific streamflow values. Moreover, specific pattern of the trend was observed to check for sensitive range of values to corresponding streamflow ranges. On the other hand, the PDPs that appeared in a straight line can mean that the specific covariate did not affect the model at all.

5.4 Forest loss effect to full-scale predicted streamflow for Research Question #3

5.4.1 Full-scale prediction

A full-scale, per subwatershed, and automated streamflow prediction from 2000 to 2016 was generated using the best prediction model out of the different grouping-learning methods. Although the accuracy measures were capable enough to show which training model method was the best, still, all models from varying *valuetable* grouping methods were run to compare the full-scale predictions. This was more applicable to the basin and size-level models since the subwatershed models were more prone to low accuracy result because of limited training power. Moreover, comparing the full-scale prediction was a good tie-breaker among models with very close results. As a rule of thumb, the model selection for full prediction only chose one grouping-learning method model and not mixed.

Summary statistics

Summary statistics of the full-scale predicted streamflow were derived in box plots to capture an overview of the predicted streamflow. The box plots were graphed per watershed size to have a better value scale distribution thus being more visually appealing. A separate plot for small and large subwatersheds were produced as well. This was helpful to provide the overview of wet and dry seasons with respect to the median value. Moreover, the plots characterized the subwatershed seasonality extremes via the range of the minimum and maximum values.

Time series graphs

Predicted streamflow and precipitation graphs or simply "hydrographs" were produced and assessed to investigate if the predicted water flows were rainfall-responsive seasonally. In other words, hydrographs were used to check whether the subwatersheds discharge higher if there is more rain. Several measures are being used in hydrology to check how rainfall and streamflow are linked. One common measure is lag time which measures the time difference between peak rainfall and peak flow. The larger the lag time means the longer the travel time of water to the discharge point. However, this measure was not used for this study because of the very high temporal resolution of the prediction as lag time are mostly used for flood modeling. The other incidental use of the hydrographs were to assess whether the subwatersheds have potential regulating infrastructures like dams since these structures alter water flows. This step further validated the subwatershed with poor prediction and low accuracy result as regulated subwatersheds behave differently especially during dry season. Theoretically, subwatersheds with dams have steady water flows even during dry season because of water regulation by purpose. On the other hand, it has lesser flows during rainy season because of water storage. Validation with the basin master plans was done for those subwatershed with dam-like behavior as indicated by the hydrographs.

The hydrographs of the 6 validation sites were created also to inspect if the predicted streamflow from subwatersheds without observed data are also capable of being rainfall-responsive. That assessment

also served as an accuracy measure.

5.4.2 Forest loss and streamflow

Deviation from historical mean

Three "well-predicted" subwatersheds and those with historical data were used to investigate how full-scale streamflow prediction deviated from historical mean flows because of limited data availability for all 21 areas. Two set of graphs were prepared for this section: (1) graph comparing with and without FLAC relative to historical mean flow, and (2) comparing dry and wet season relative to historical flow data. Seasonality was captured by accustoming to the climate type of the area. This was automated also using functions in R (*aggregate, melt, filter*) to summarize monthly flows and arrange accordingly per different subwatershed season. The historical seasonal flow was drawn as a threshold line to that of the predicted seasonal flow averages. Its deviation was closely investigated per season to assess the potential impacts of forest loss. Anomaly can be assessed by comparing predicted streamflow in a given month compares to its long-term average (Shortridge et al., 2016). Standard deviation of the historical mean was used to declare an extreme flow as an outlier.

Mean daily streamflow per season

Yearly mean daily seasonal flows per year were graphed and regressed with forest loss for 6 selected subwatersheds as well as to 6 validation sites. This was one of the key indicators on how forest loss affected the flows seasonally. The choice of subwatersheds were those that performed well as per accuracy measures in section 4.2.4 and rainfall-responsiveness. The purpose of this step was not to find good regression results because the data points were so limited (16 points/subwatershed). The revealed pattern (inverse or not) of forest loss and seasonal flows was more favored. However, R^2 was still reported.

The regression for wet season produced one collated graph since all subwatersheds depicted a positive linear relationship between forest loss and wet flows.

The effect of precipitation was evaluated also using the same regression procedure since rainfall can also affect the seasonal streamflow pattern. Moreover, this step served as the basis which drives the seasonal trend better - forest loss or precipitation?

Rainfall - peak flow ratio (RPFR)

Peak flows of extreme rainy days were divided into its corresponding rainfall to get the ratio of both as a relative measure on how peak flows change over time. The lower the RPFR value the more streamflow it generates relative to rainfall. The basis for determining "extreme" was sorting the rainfall value and filtering the top 100 rainy days over the 16-year time series. This step served as an indicator to assess the change of forest regulation service in wet season. The most rainfall-responsive subwatersheds were chosen as test sites.

The yearly (2000-2016) rainfall-peak flow ratio of the selected subwatersheds were regressed again with forest loss to assess how strong the correlation is.

Scenarios

A forecasting method was generated using the linear relationship of forest loss to dry season flow and RPFR from unregulated subwatersheds. Three scenarios based on forest loss rate were created: (1) current rate, (2) additional 5% rate, and (3) additional 10% rate. The unit of measure was relative (% increase/decrease). Only those subwatersheds with above 0.30 in R^2 were modeled.

6 Results

6.1 Predicted flow assessment

This subsection assesses the predicted/simulated streamflows from different grouping-learning methods relative to the observed/actual data. Accuracy measures (R^2 and NSE) comparison and in-depth predicted streamflow assessments are the gist of this section.

6.1.1 General findings

Accuracy results using held-out or test data varied per RF grouping-learning method. One general finding was that grouping the *valuetables* instead of having 1:1 *valuetable*-model yielded better accuracy results. In other words, subwatershed-level grouping performed beyond the other grouping methods. This finding was evident when mean accuracy (R^2 and NSE) result is the basis as the "learn-all" method seemed to outperform the other grouping methods as shown in Figure 7. The results of model testing (10% of remaining observed data) for basin and subwatershed levels are shown in Annex D.

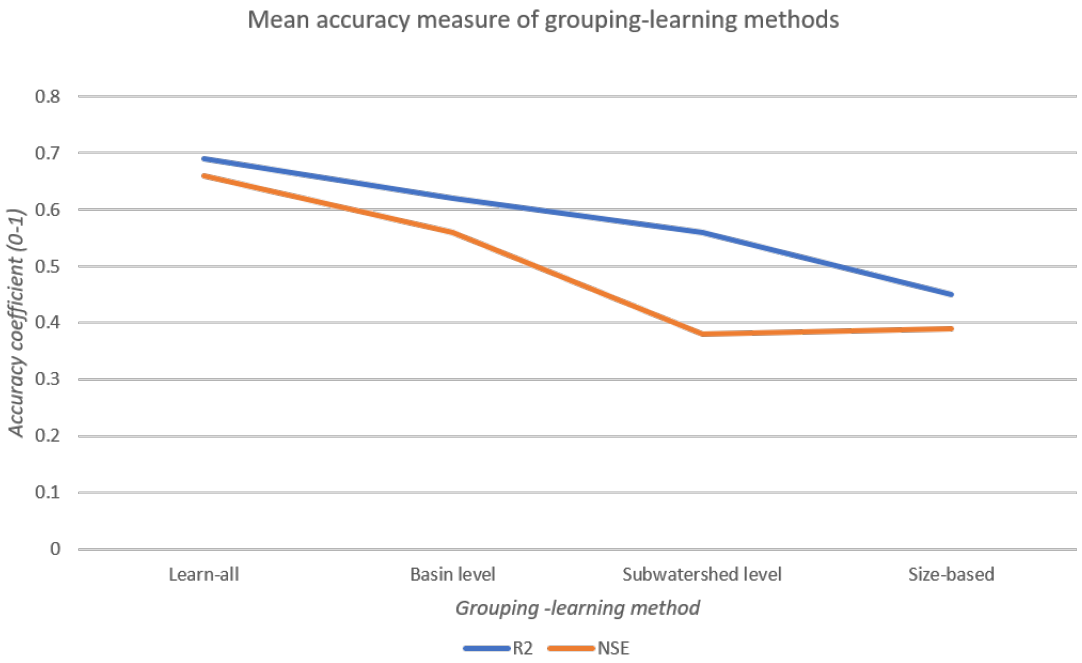


Figure 7: Mean accuracy measures from different random forest grouping-learning models.

Another common finding to all grouping-learning method was that high absolute number of observations was a non-factor to attain higher prediction accuracy, again using R^2 and NSE as measures. This finding was more evident from the basin and subwatershed grouping-learning methods where models are comparable because the methods have 6 and 21 sub-models, respectively. Similarly, accuracy indicator R^2 was not correlated with watershed size. Shown in Annex C is the relationship between the R^2 and the said two indicators (number of observations and watershed size). The said two observations hint that model training accuracy can be affected more by the combination of watershed features itself and not by one single factor especially the number of observations. Lastly, a general observation

was that R^2 and NSE coincide with a 0.94 correlation value. However, NSE was always lower than R^2.

6.1.2 Per grouping-learning method findings

The "learn all" method, which collated individual *valuetables* into one very informative table, yielded a significant accuracy reulst of 0.69 R^2. The regression results of model testing (10% of over-all data) is shown in Figure 8. This result can mean that training the model with every bit of data available can result to the highest model accuracy even though watersheds are from different basins. This finding further means that the model is able to nullify the extreme low R^2 observed from the subwatershed models. In other words, those irrelevant information at subwatershed-level can make more sense when combined with other subwatershed information. The last observation was NSE result just deviated a bit at 0.67 for learn-all method.

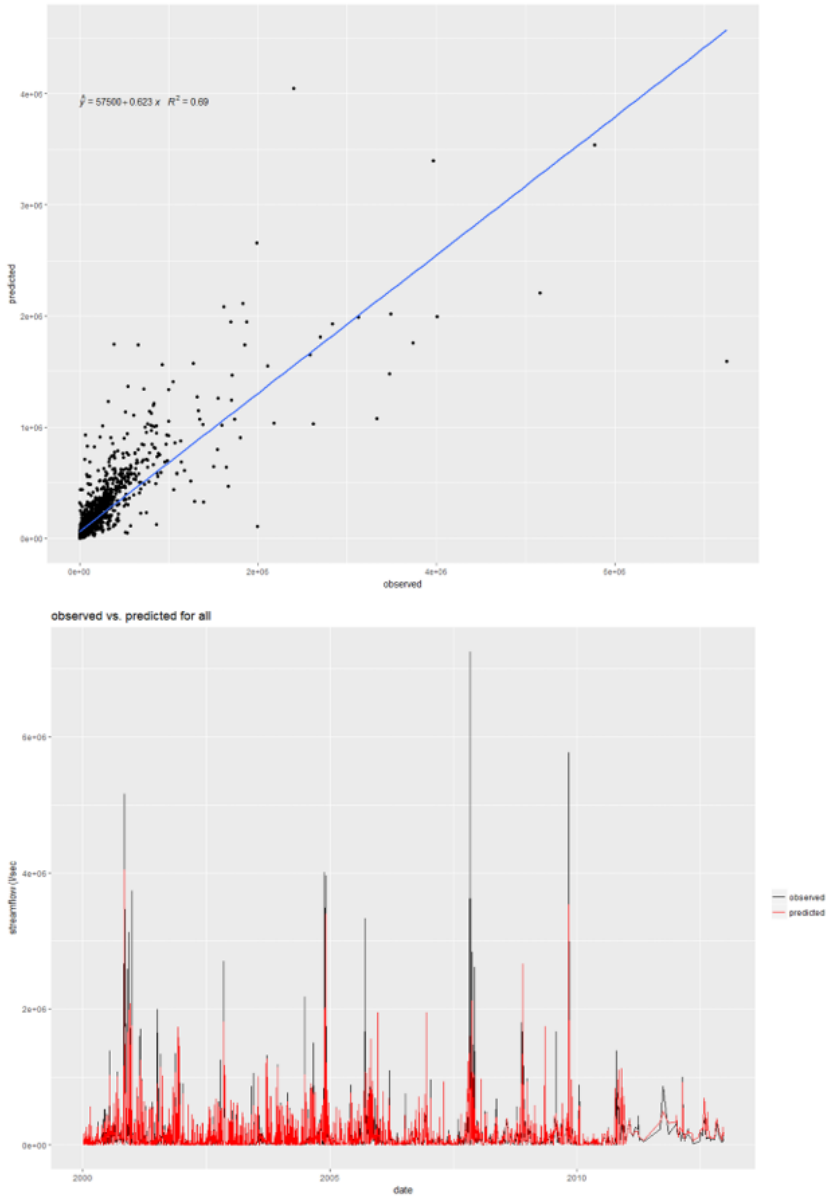


Figure 8: Model test result using 10% of remaining observed data for âĂIJlearn-allâĂİ grouping-training method.

Basin models performed relatively good as only one model (abrb) was below the significant threshold of R^2 (> 0.60). There was no clear trend relating R^2 to sample and basin size. The basin with most number of sample size (crb) was only 4th in correlation results ranking. However, basin abrb, being lowest from the correlation results, was also the smallest in area and sample size. Judging the observation-accuracy for basin abrb can be a combination of the earlier said findings that models from small subwatersheds and ungrouped *valuetables* perform relatively poorer as both situations are present in basin abrb. On the other hand, the basin (arb) with the highest R^2 and NSE did not have clear connection with the sample size. Again, it signifies that the number of training data is not the mere basis of attaining high accuracy result. The per basin accuracy results, sample size, and area are summarized in Table 7 and visualized in Figure 9 below.

Table 7: Accuracy results, sample size, and basin area.

Basin	Sample size	Area (ha)	R^2	NSE
aarb	6,954	25,5396	0.59	0.58
abrb	3,461	1,461	0.44	0.30
arb	5,113	68,116	0.80	0.73
crb	27,402	4,120,698	0.6	0.58
mrb	2,191	63,444	0.65	0.64
prb	8,569	977,516	0.67	0.5

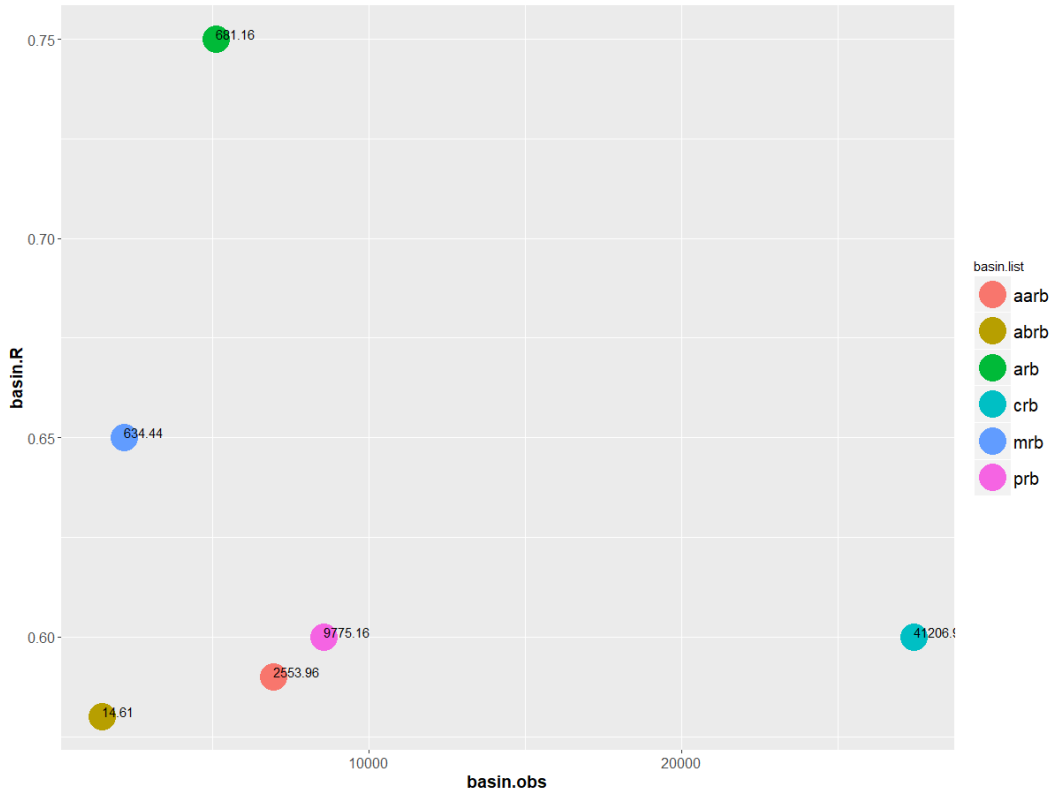


Figure 9: Summarized plot of accuracy result, number of observations, with labels of watershed size for basin models.

Higher variability of accuracy outcome was observed for the subwatershed models as shown in Table 8 judging from the results of R^2 and NSE. The variability can be attributed to the method itself since subwatershed-level grouping-learning method essentially took away the essence of static covariates because no *valuetable* grouping happened. Almost half (48%) or 10 out of 21 had positive strong correlation results or those above 0.6 R^2. Bulk of the relatively poor models belonged to basin crb. Comparing it with the basin-level learning, the 1:1 method performed way below. For example, those belonging to basin crb resulted to a mean of only 0.41 R^2 or beyond by 0.19 compared to crb at the basin-level method. This discrepancy was also evident, but in lower magnitude, in basin prb where 1:1 subwatershed models were beyond in R^2 by 0.05. Moreover, sample size with an average of 2,461 observations and subwatershed area were also not affecting the accuracy results but here, accuracy-sample size variability was more evident than the basin-level results. Affirming variability are the top left points in Figure 10 where basins showed significant accuracy results given a lower sample size. Looking at the climate type, those belonging to climate type 1 or pronounced wet and dry season had better accuracy results which can hint that distinct seasonality can be learned by the random forest algorithm well.

Table 8: Accuracy results, sample size, and subwatershed area.

Subwatershed	Area (ha)	climate type	Sample size	R^2	NSE
aarb_a	236,119	3	3969	0.56	0.55
aarb_n	19,277	3	2985	0.72	0.71
abrb_s	1,461	1	1461	0.44	0.30
arb_b	31,609	1	2191	0.82	0.81
arb_c	36,507	1	2922	0.1	-1.5
crb_a	104,525	4	3287	0.62	0.6
crb_be	27,211	3	1645	0.21	0.2
crb_bu	1,851,853	3, 4	2040	0.65	0.58
crb_d	169,590	4	3584	0.4	0.38
crb_j	293,539	4	4384	0.62	0.58
crb_m	215,601	3	1096	0.51	0.27
crb_p	1,187,380	4	3409	0.65	0.59
crb_s	85,267	3	2843	0.2	0.11
crb_t	66,994	4	3653	0.1	0.019
crb_u	118,738	4	1461	0.18	0.16
mrb_s	63,445	1	2191	0.65	0.65
prb_a	622,280	1	2026	0.2	0.17
prb_b	37,406	1	1826	0.76	0.68
prb_c	84,287	1,3	912	0.86	0.72
prb_p	85,739	1	1432	0.4	0.39
prb_r	147,804	1	2373	0.91	0.91

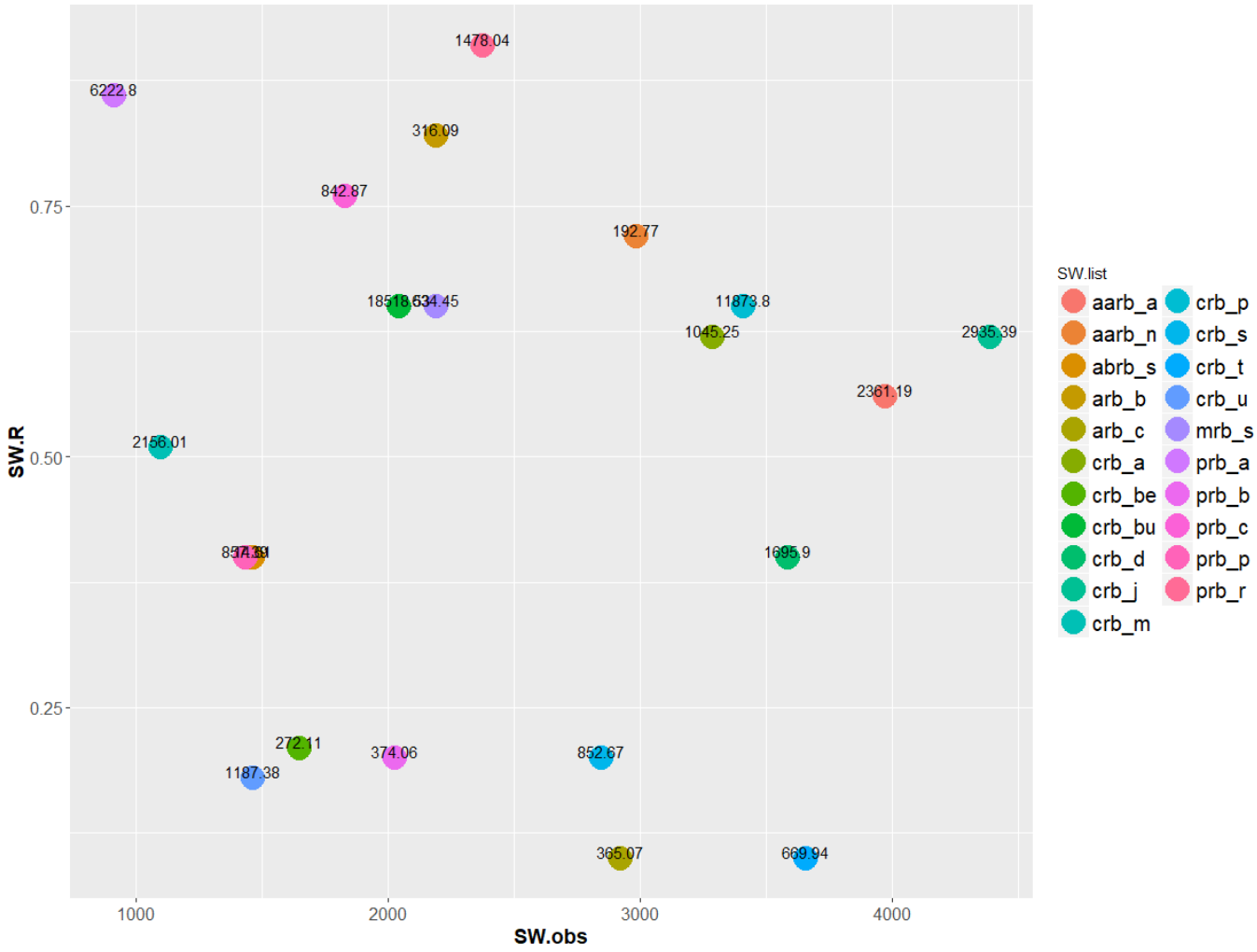


Figure 10: Summarized plot of accuracy result, number of observations, with labels of watershed size for subwatershed models.

The last grouping-learning method size-based grouping yielded 0.35 R^2 for small subwatersheds, 0.56 R^2 for large subwatersheds using 24,061 and 27,629 number of observations, respectively. The 0.21 deviation among the two sub-models is considered significant and essentially translates that large subwatershed information can be more sensible to the RF algorithm. Moreover, the smaller subwatersheds are just "part" of the larger ones in the case of this study and it seems the smaller subwatersheds exhibit more heterogeneity.

6.1.3 In-depth comparison of simulated and observed flows

The graphs in (Figure 11) showcase the daily average flow per month deviation of the simulated data from the actual data at basin level. Generally, the predictions deviated more during rainy months as there was an average of 3.8 rainy months in a year with over-predicted steamflow. However, there were some months that under-predicted rainy months but are just 1.3 month per year on the average. The rainy season over-prediction can be observed also from the regression graph in (Figure 8) where bulk of points were above the 1:1 line at 5e+05 to 6e+06 range. On the other hand, the predicted streamflow during dry season coincided better with the observed data better than that of the rainy months.

The seasonality became more obvious as line graph smoothing was observed when the daily flows

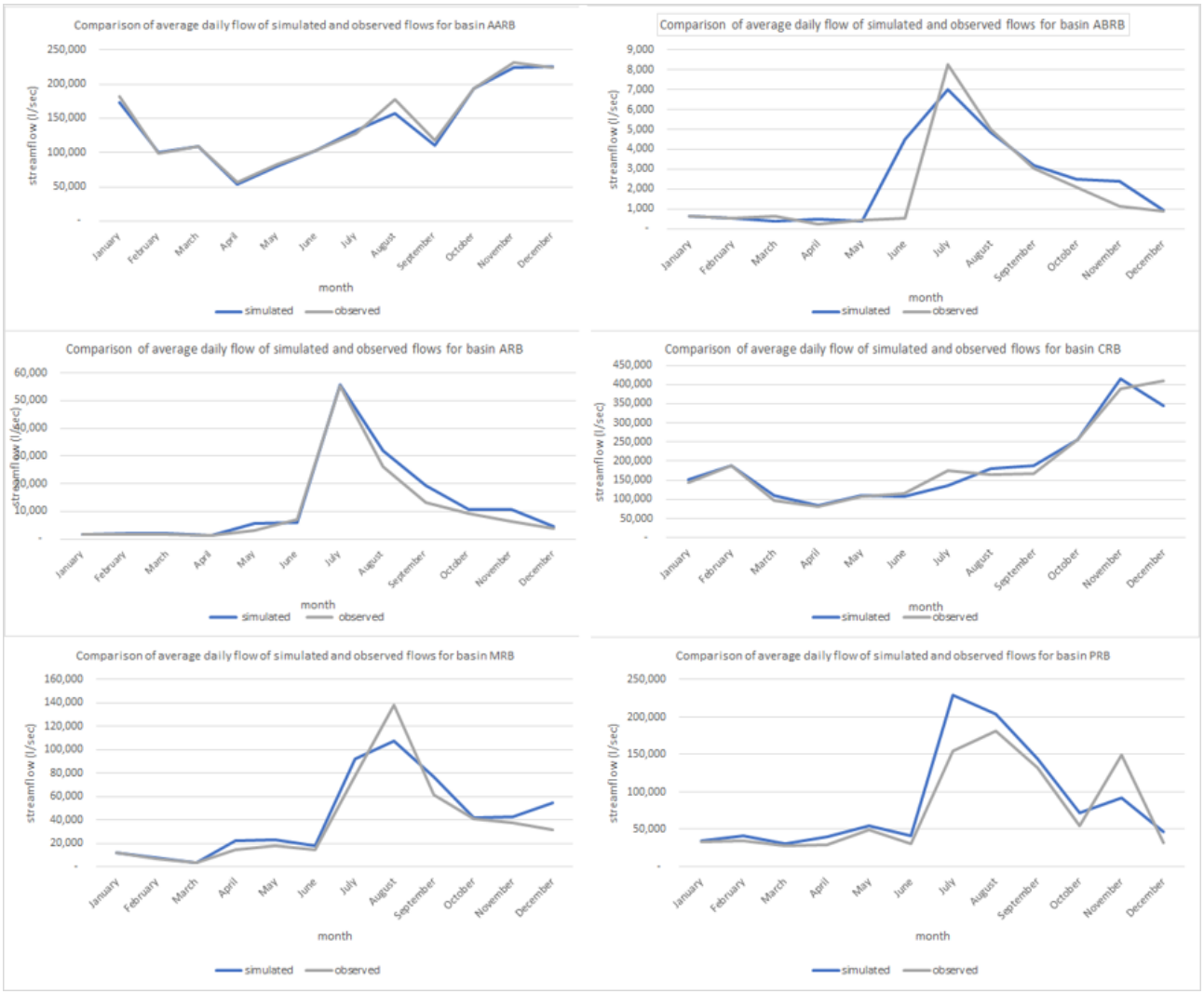


Figure 11: Predicted daily mean streamflow per month using test data compared with observed data for 6 basins.

for both simulated and observed are averaged in a monthly basis. There were no evidences of the extreme flows effect which can be attributed to the "averaging out" effect or cancellation of highs and lows that mixes extreme flows to more regular flows.

The prediction using the test data came up short in predicting peak flows of severe rainy days in general as observed in Figure 8 above. This was also the observation from the three other grouping methods seen in the Annex F. To reiterate, the remedy during the midway of the thesis was to sample a good performing subwatershed to include a maximum daily rainfall covariate. Incidentally, the comparison of with and without max rainfall in a day revealed that indeed predicted peak flows can improve (Figure 12) by as much as 30%. It can be seen also that the flows aside from the peak or those very high discharge days remain similarly the same.

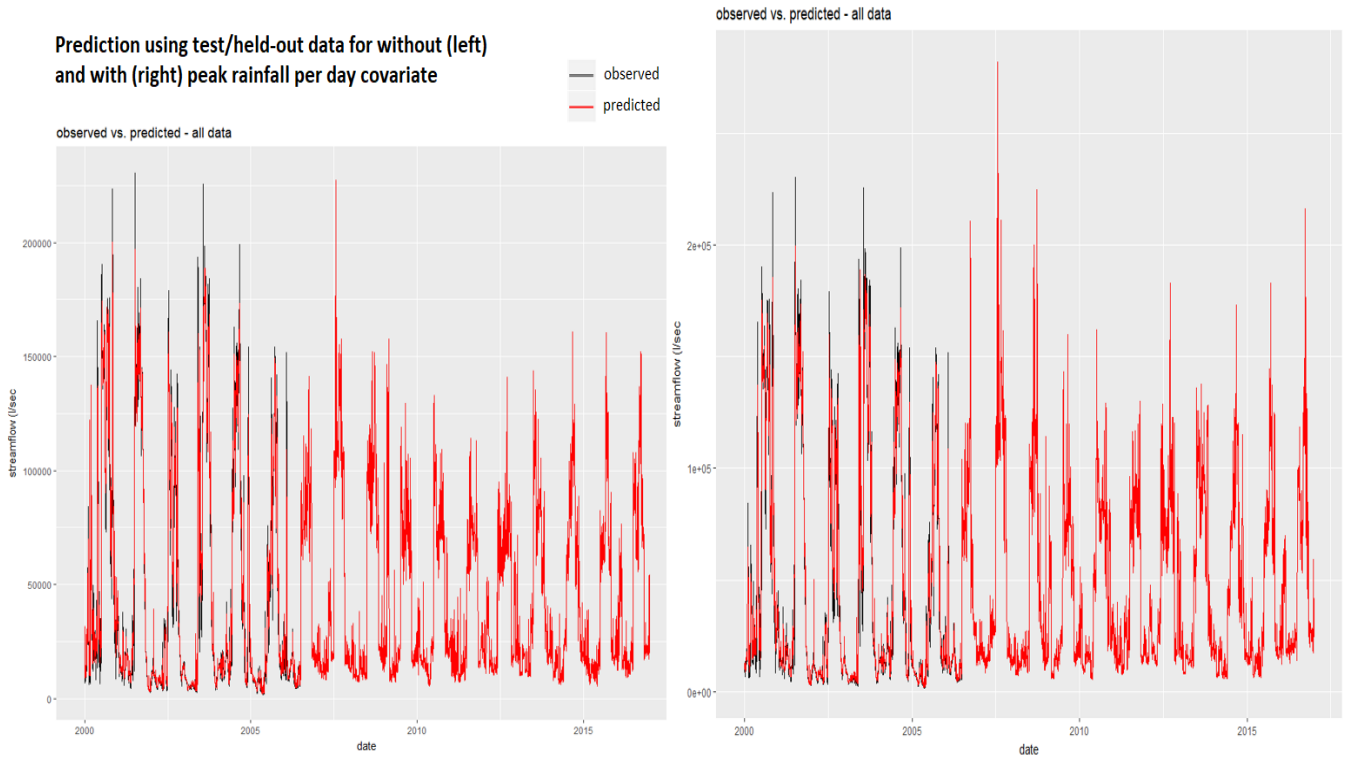


Figure 12: With (left) and without (right) peak rainfall per day covariate full scale prediction.

6.1.4 Comparing with other models

The results for basins mrb and aarb were compared with other hydrologic models from basin master plans (aarb and mrb) and the study by Rawlins et al., 2017 for mrb. It appeared that the monthly average predicted streamflow by this study exhibited similarity between other models especially for mrb as shown in (Figure 13). However, in basin aarb, the two model results were not that very similar (but still linearly related) mainly because of the temporal scale discrepancy as the other model used 1980 to 2012 range.

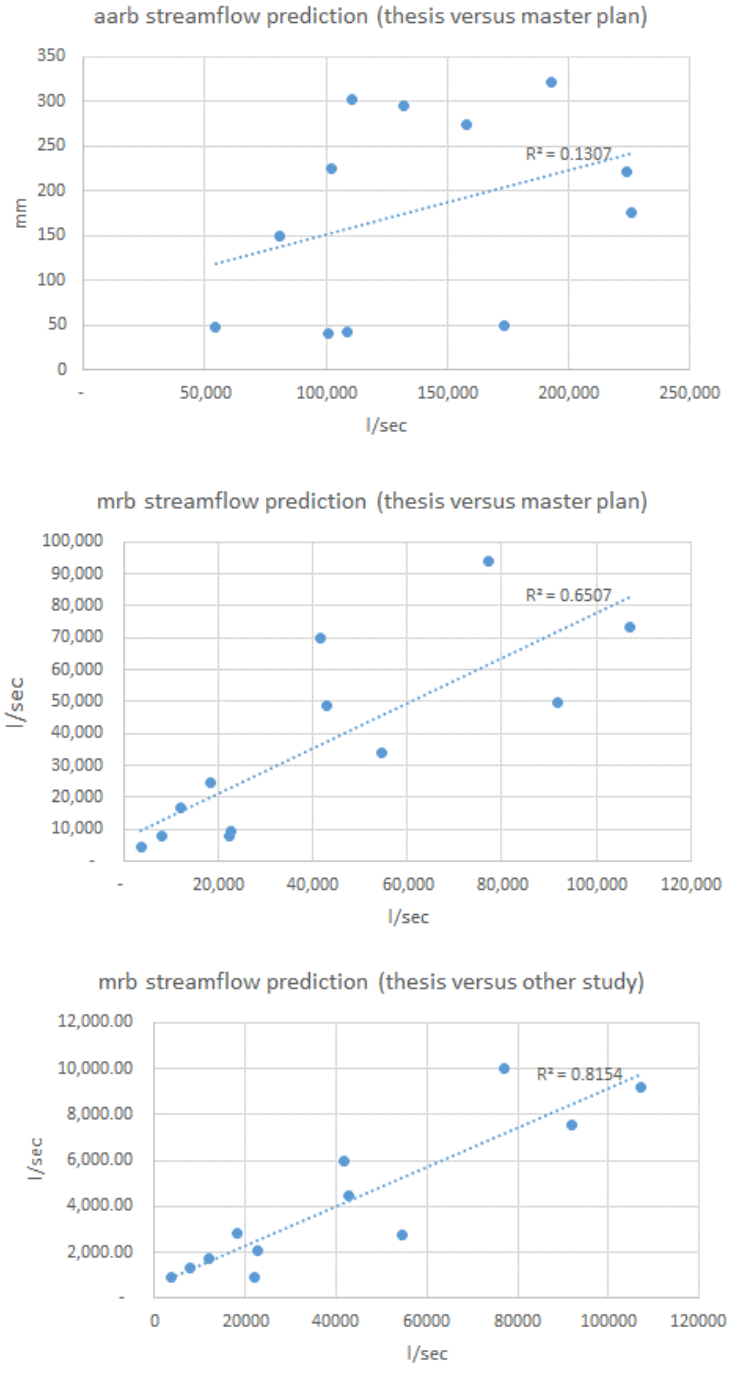


Figure 13: Regression graphs of thesis results versus other hydrologic model results from master plans and other studies.

6.1.5 Validation sites

All validation sites depicted seasonality based on the full-scale prediction using the learn-all RF model as shown in Figure 14. The prediction also coincided with the climate type of the sites where distinct seasonality was observed in aarb_vs, abrb_vl, abrb_vs belonging to climate type 1 (pronounced wet and dry season).

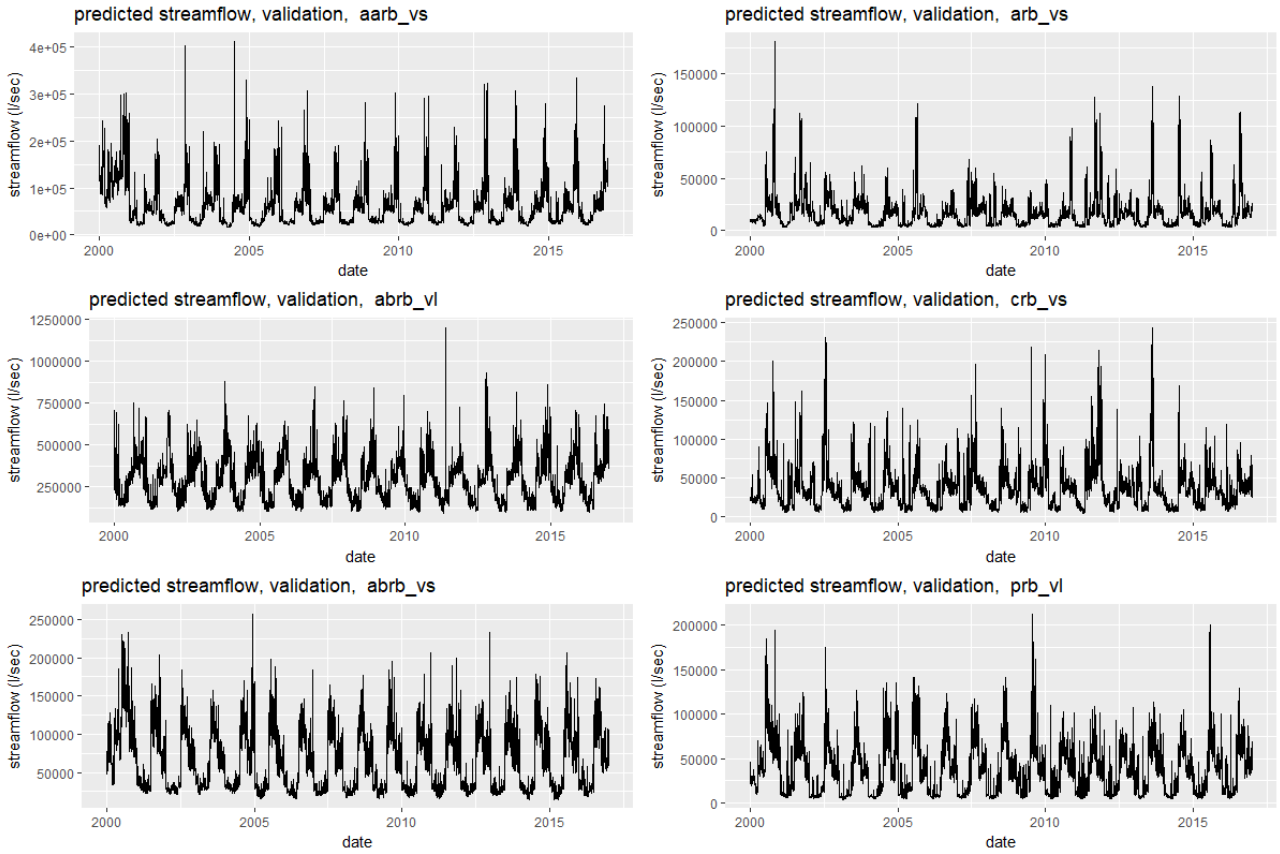


Figure 14: Full-scale streamflow prediction for 6 validation sites.

6.2 Forest loss and associated covariates (FLAC) assessment

This subsection pertains to the assessment of the effects of FLAC covariates to the random forest models and predictions using 3 measures (permutation, removal, and partial dependence plots) for each of the grouping-learning method.

6.2.1 Permutation indicator

The *randomForest* function produced an importance measure based on covariate permutation called as "variable importance measure" or simply VIM. The measure VIM was derived using the average difference of accuracy decrease when a covariate of interest is to be permuted. The FLAC, when permuted, affected the model by decreasing accuracy of 18% on the average. This effect can be attributed to the yearly change in forest loss values resulting into cumulative decline in forest cover as the RF model seems to be reactive to this trend which is reflected also, of course, to its associated covariates. The FLAC impact as per VIM can be observed visually for the "learn all" model shown in Figure 15 and the rest of the models in Annex E. Highest among the FLAC covariates at 20% accuracy decrease was forest loss and forest cover which are essentially forest conversion outcome. A bit below (12%) the importance measures were associated covariates Curve Number and Manning's coefficient. The result for these two associated covariates makes sense since the two are relative values with narrower range of values compared to absolute values in hectares which can be more sensitive to permutation.

The FLAC were within the top 13 (out of 58 covariates) important to the basin models and contribute for 18% of the total VIM (Figure 16) for all basins on the average. Majority of the VIM contribution were given by weather covariates due to its higher variability in nature. Note that these are daily data and when permuted, can affect the model performance worse compared to yearly dynamic covariates which FLACs are. Covariate month (mo) was the top VIM contributor for the 6 basins or simply the "most important" covariate indicating the distinct seasonality of Philippine climate. Interestingly, covariate month has lesser temporal resolution than the weather covariates (monthly versus daily), nevertheless, has higher VIM value.

The pattern on how FLAC vary per basin was noticeable. In huge basins like prb and crb, the amount of forest loss was relatively bigger also. That magnitude seems to affect VIM scores when absolute forest loss in hectares are to be permuted. For crb, the conversion effect was amplified by the high importance values of forest cover. This basin also has the highest forest loss rate among all basins caused by rampant illegal logging in early 2000s. Moreover, years 2000-2004 for crb had the highest forest loss incidence among other basins which resulted into a steep increase of forest loss graph from that phase. Unfortunately, there is no temporal measure given by the VIM to better support the influence of the abrupt forest loss change within 2000-2004. For prb, covariate *LC.for* or forest cover was the second highest among all covariates. Incidentally, that basin despite having low forest cover, but, with considerable forest loss yearly at 4% seems to trigger the model. In other words, it is equivalent to higher cover-loss ratio also which makes forest cover a significant VIM contributor. Each VIM graph per basin is shown in Annex E.

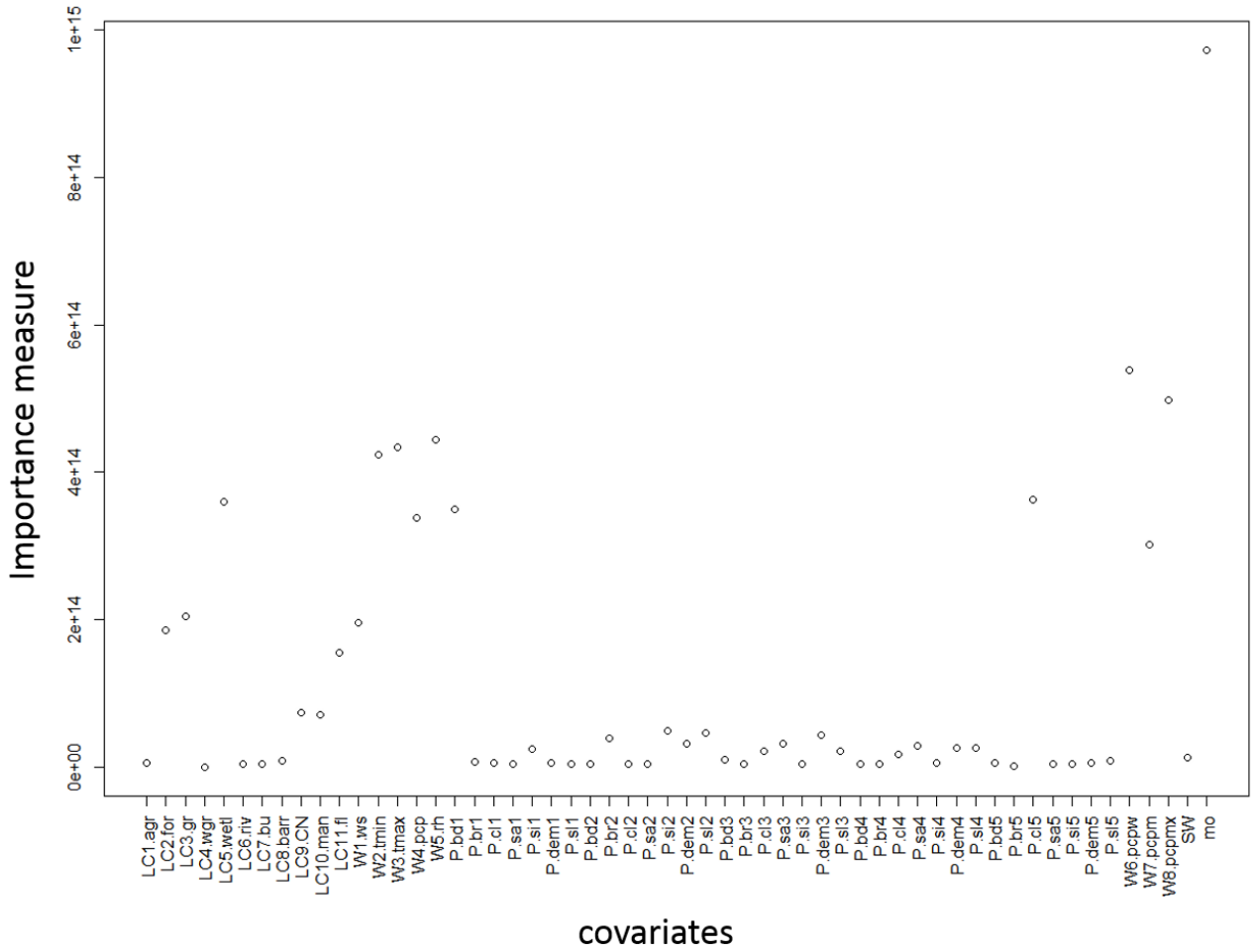


Figure 15: Per covariate importance measures for learn-all prediction model.

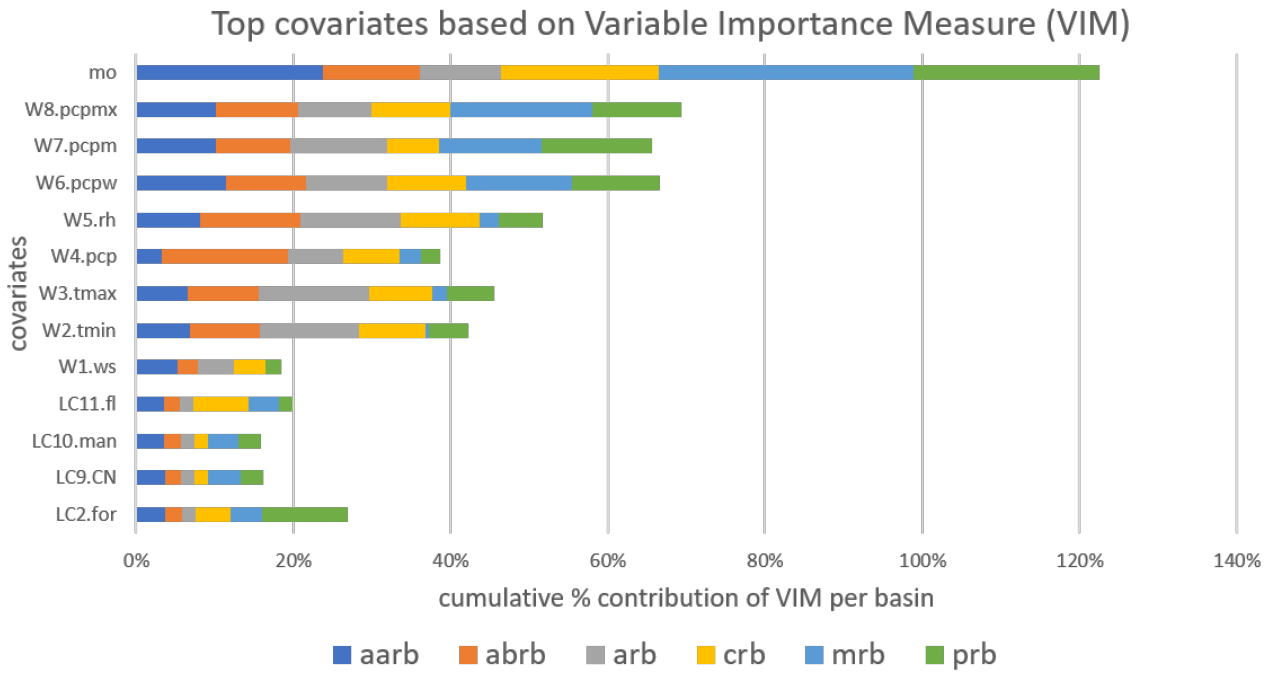


Figure 16: Top covariates with %breakdown of variable importance measure per basin.

Variable importance measures of *randomForest* differed by that of *ranger* although both are packages for random forest algorithm. Again, this variation happened since the ranger script was not operational at subwatershed level due to some function-calling limitation. However, both measures (unexplained variance increase for ranger and MSE increase for randomForest) are relative measures, as such, the subwatersheds can be compared accordingly. Even at per subwatershed learning, FLAC increased %MSE if permuted. In fact, the FLAC at the subwatershed-level VIMs had more model effect as FLAC were consistently within the top 10 contributors of VIM for all subwatersheds. Same observation with the basin-level VIM was observed as forest loss and forest cover covariates were the top VIM contributor among the FLAC group. Highest influence of the said two FLACs was observed in prb_r which could affect MSE by as much as 50%. This subwatershed coincidentally performed well via accuracy indicators meaning there can be a correlation between significant VIM result of FLAC and model accuracy. Another observation with the subwatershed-level VIM graphs was that FLACs tend to outperform some weather covariates. Located in Annex E are VIM graphs per subwatershed.

6.2.2 Removal measures

There was an average of 0.06 difference with a relative decrease of 17% in R^2 and 7.5E+08 change in MSE for all basins when FLAC were removed as covariates as shown in Table 9. Note that, in this case, MSE was added as a measure even though it is an absolute measure since it was easily accessed from the random forest models. To make the measure relative, percent change was computed and it turned out that FLAC increased the MSE by 4%. The discrepancy after FLAC removal based on MSE and R^2 can still be considered important even the changes are relatively insignificant (aside from basin prb with the only basin with $>$ 0.1 R^2 decrease). It is because the baseline regression results were not very high as only 3 out of 6 basins were below the 0.6 significance threshold after FLAC removal. Basins prb and arb had the highest R^2 decline when FLAC were removed which can be attributed to some abnormal high flows even in dry season and extreme flows that can be considered as outliers (Figure 17). Looking beyond the R^2 difference from with and without FLAC models, the significant decline when FLAC are removed can hint about the regulating effect of FLAC in water flows to control abnormal high flows. In mrb, it was more evident and could be affected by relatively smaller training data. Basin abrb with the smallest training data number had the highest discrepancy on R^2 with and without FLAC. Regression graphs per basin with and without FLAC are shown in Annex F.

Table 9: Effect of FLAC removal on R^2 and MSE per basin.

Basin	Sample size	R^2 decrease	% change in R^2	MSE difference
aarb	6,259	0.07	20%	2.50E+07
abrb	1,315	0.05	14%	1.00E+07
arb	4,602	0.08	23%	3.00E+08
crb	24,662	0.01	3%	3.20E+09
mrb	1,972	0.04	11%	1.20E+08
prb	7,713	0.11	31%	8.00E+08

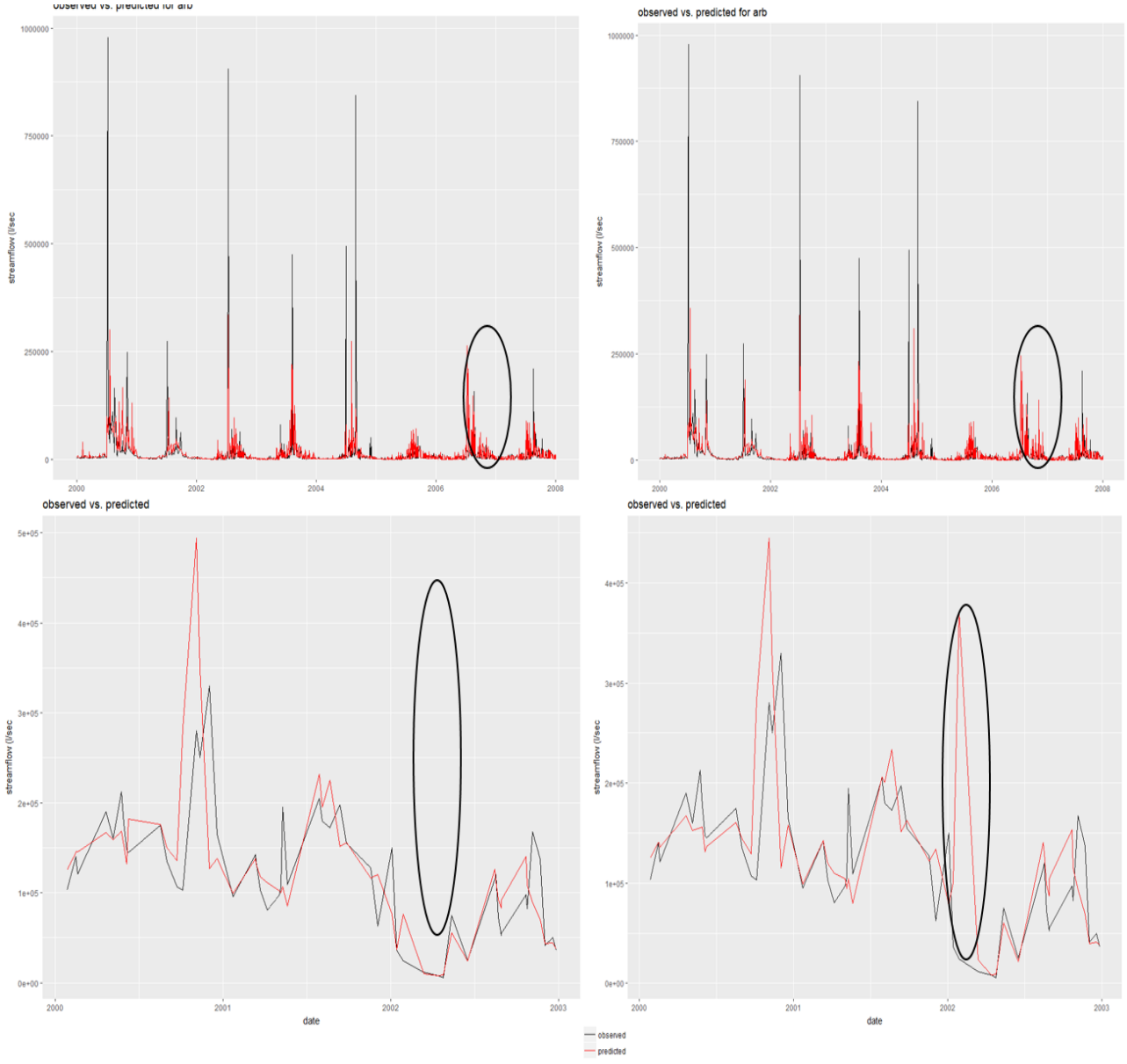


Figure 17: Outliers observed in sample basins (arb and mrb) when FLACs are removed as covariates.

A summary of the simulated and actual flows seasonal comparison (basin-level) in relative values is shown in Table 10. In an overview, both (with and without FLAC) simulated flows exhibited over-prediction of 10% on the average from the 6 basins. Dry month flows (6%) had relatively lesser over-prediction than wet month flows (14%). Seasonal comparison of with and without FLAC predicted flows revealed that the latter under-predicted streamflow during dry season in 4 basins while it over-predicted during wet season for 5 basins compared to the former. This simply means that without FLAC, the seasonal flows can be a bit more erratic than the models with FLAC.

Table 10: With and without FLAC predicted flows % change compared to observed flows seasonally

Basin	with FLAC		without FLAC	
	dry	wet	dry	wet
aarb	-2	-2	-10	7
abrb	20	20	-10	24
arb	4	23	2	29
crb	4	3	4	4
mrb	10	7	13	-3
prb	18	23	12	25

6.2.3 Partial Dependence Plots (PDP)

In the case of the 21 subwatershed RF models, 20 out of 21 models depicted that FLAC influence the streamflow prediction by having linear trend represented by Figure 18, the rest of the graphs are located in Annex F. This is a good indication since a mere straight line graph PDP means the covariate does not influence the dependent variable at all (Strobl, 2008). In other words, a sloping PDP line is essentially equivalent to a strong streamflow prediction link. To be more specific, the linear observation from PDPs, either inverted or not, exhibits that there are low-high/high-low and high-high/low-low values link between streamflow and FLAC.

The trends shown in Figure 18 for the 4 FLAC was common to 70% of the models where a positive linear trend was observed between streamflow to forest loss and Curve Number. That two sub-graphs above show that low to high forest loss and Curve Number values have strong influence in predicting low to high streamflow, respectively. This result make sense since forest loss are accumulated and higher Curve Number meant more deforestation.

Oppositely, the two sub-graphs below in Figure 18 show that forest cover and Manning's coefficient were strongly affecting streamflow prediction the other way around compared to forest loss and Curve Number. The finding can be translated to a high-low/low-high link between streamflow and the said two FLAC. This pattern is essentially a result of forest cover covariate declining constantly per year as well as Manning's coefficient since its value declines when deforestation happens.

The PDP graphs revealed some pattern also relative on how steep and gradual the drops or rise are. This can be attributed on how a particular FLAC covariate values influence per certain streamflow ranges. Steep drop and rise were observed in Manning's and Curve Number sub-graphs respectively (LC10.man and LC9.CN). It can mean that wider streamflow range can be influenced by narrower covariate values. On the other hand, the gradual drop and rise observed from forest cover and loss (LC2.for and LC11.fl, respectively) had more balance between the magnitude of the pairing values. The gradual drop or rise of the line is capturing seasonality of streamflow better essentially. In other

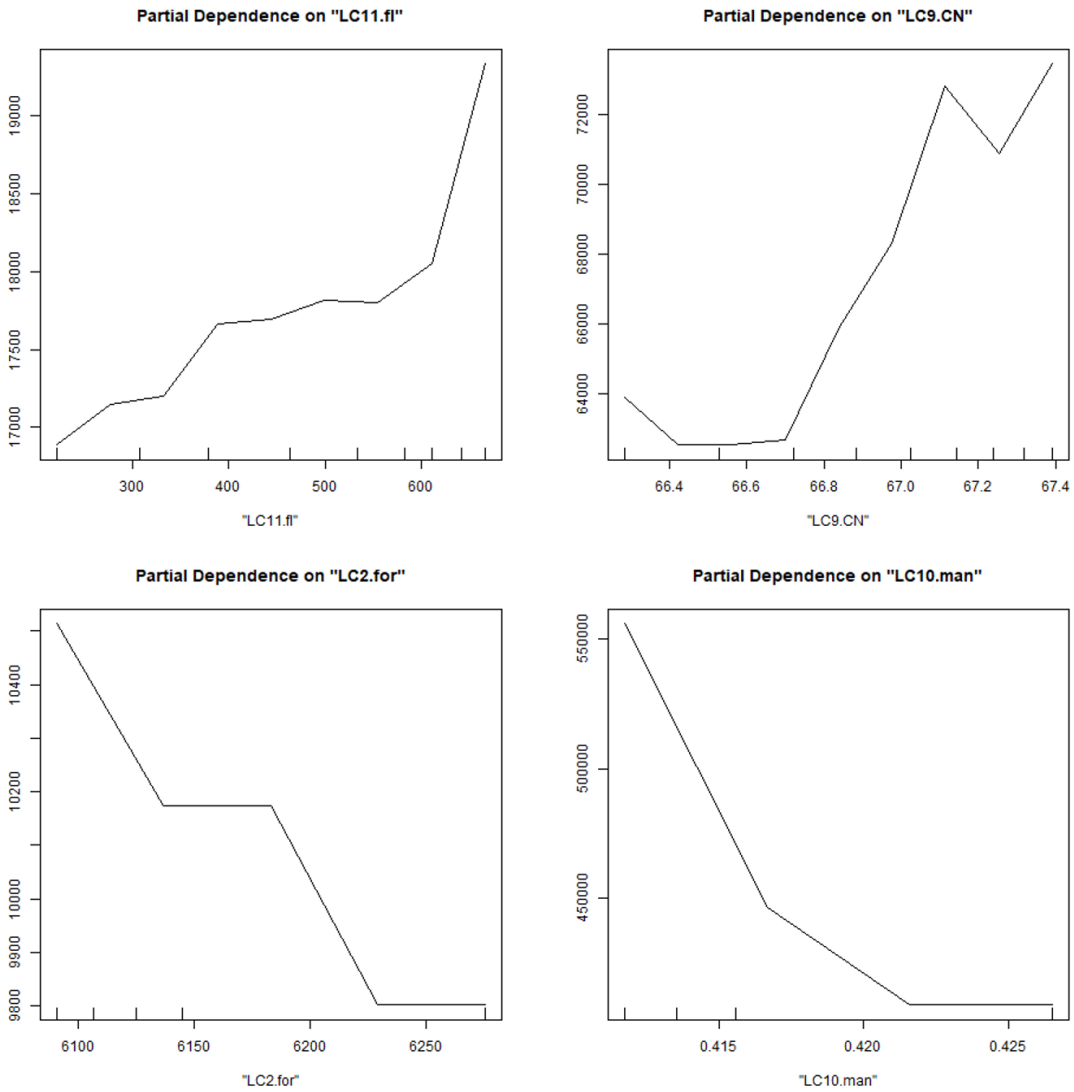


Figure 18: PDP graphs of selected basin for the 4 FLAC (top left: forest loss, top right: Curve Number, bottom left: forest cover, and bottom right: Manning's coefficient.

words, FLAC with absolute values has more "magnitude effect" than relative ones. Note that the dependent variable streamflow has absolute values also. As such, the finding in this last paragraph can be interpreted also as absolute-absolute variable had sensible prediction impact than absolute-relative variable pairing.

6.3 Forest loss and full-scale predicted streamflow

Overview

This section shows 2 main results about the full-scale (2000-2016) predicted streamflow and the effect of forest loss to it. First, 3 indicators were used to evaluate the full-scale predicted streamflows. These are box plots, hydrographs, and start-end year comparison. Secondly, the effect of forest loss was evaluated via 2 indicators: the deviation to historical mean daily streamflow; and the regression of forest loss to dry flows and RPFR. The summary of the key results for this section is shown in Annex G. The summary table also shows supplemental information used for highlighting results in this section like number of water regulating structures, model accuracy, climate type, and size.

6.3.1 Full-scale prediction assessment

Box plots

The summary statistics in the form of box plots for all subwatersheds together with those assigned as small and large are shown in Figure 19. The proportion of boxes in the figure reveals that most subwatersheds had longer wet season in general having larger values above the median line. Moreover, the range of the upper values were also way higher as observed from the dashed line. This simply depicts the actual climate type in the Philippines which has relatively more rainy months than dry months.

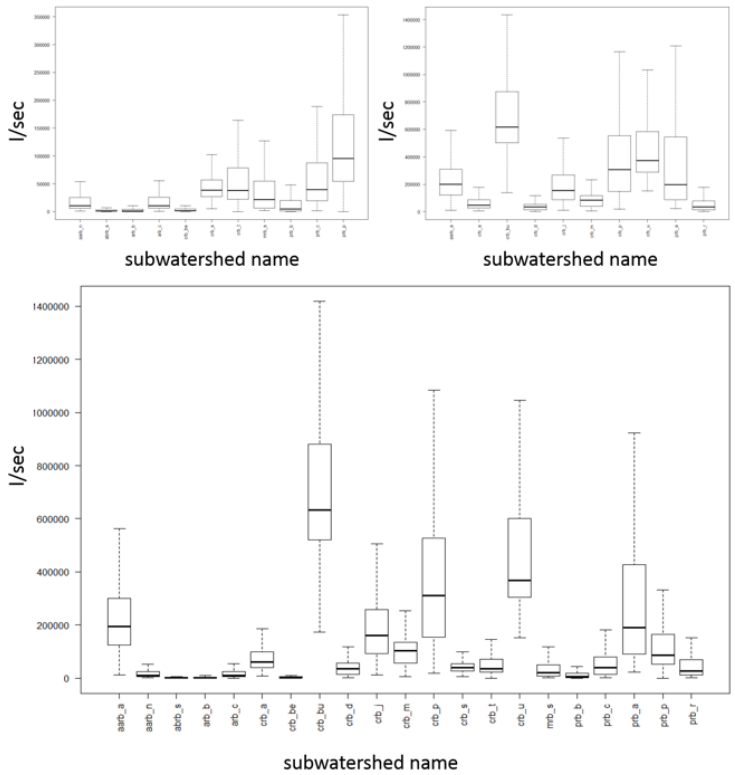


Figure 19: Box plots of full-scale predicted streamflow for small (top left), large (top right), and every (below) subwatersheds.

Hydrographs

Hydrographs or streamflow-precipitation graphs served as the second indicator to assess if the full-scale streamflow predictions are good. To reiterate, rainfall-responsiveness is defined as the linear reaction of streamflow discharge to a specific rainfall event i.e. high rainfall is equals to high streamflow. Using the "learn all" grouping-learning method, the full-scale prediction of 13 out of 21 subwatersheds looked to be rainfall-responsive. These were aarb_a, aarb_n, abrb_s, arb_b, crb_a, crb_bu, crb_d, crb_j, crb_p, mrb_s, prb_b, prb_c, and prb_r. Figure 20 displays the most rainfall responsive hydrographs, the rest are shown in Annex H. Judging visually, the four sub-graphs in Figure 20 have a distinct seasonal flow throughout the 16-year stretch as streamflow corresponded to the amount of daily rain in most cases. Incidentally, most of the rainfall-responsive subwatersheds were also the subwatersheds with high accuracy prediction (0.6 and above R^2) when individually trained. In other words, the full-scale prediction agreed well with the accuracy results. Moreover, it can indicate that the characteristics of these areas are so definitive as both the "learn all" and individual models can predict the flows similarly. On the average, rainfall-responsive subwatersheds had 0.66 in R^2. More interestingly, the rainfall-responsive subwatersheds were those considered to have the least water regulating infrastructures as shown in Annex I.

The subwatersheds used for validation all depicted rainfall-responsiveness. It happened that these 6 subwatersheds were pre-selected not to have regulating structures. The graphs can be seen in Annex J.

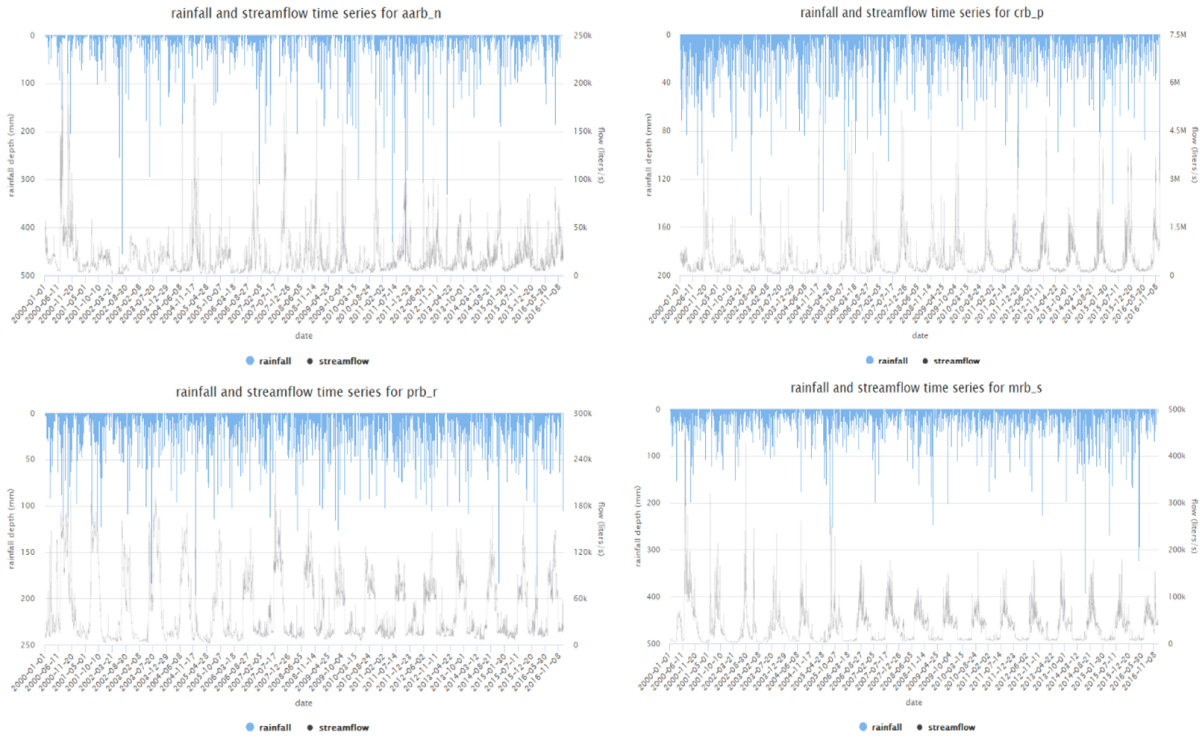


Figure 20: Subwatersheds representing the rainfall-responsive group.

On the other hand, there were 8 out of 21 subwatersheds that did not depict rainfall-responsiveness. These were arb_c, crb_be, crb_m, crb_s, crb_t, crb_u, prb_a, and prb_p. Least rainfall-responsive subwatersheds were mostly from crb basin as shown in Figure 21. Basin crb has a lot of big and small dams constructed for irrigation and flood control (Cagayan River Basin Master Plan, 2013) as shown in Annex I. The hydrographs that were not rainfall responsive were also subwatersheds that yielded low accuracy having just an average of 0.23 in R^2. Judging from the hydrographs, it can

be observed that the time series streamflow pattern was a bit irregular despite showing a seasonal trend. Looking at specific season and year revealed some instances. In crb_t for example, water flows tended to become more regulated from mid-2007 onwards. A similar observation was seen in the crb_m graph where rainfall responsiveness was observed until year 2003, beyond that year were more regulated flows even in rainy season. These flow irregularities can be caused by the start of the water regulating structures operation.

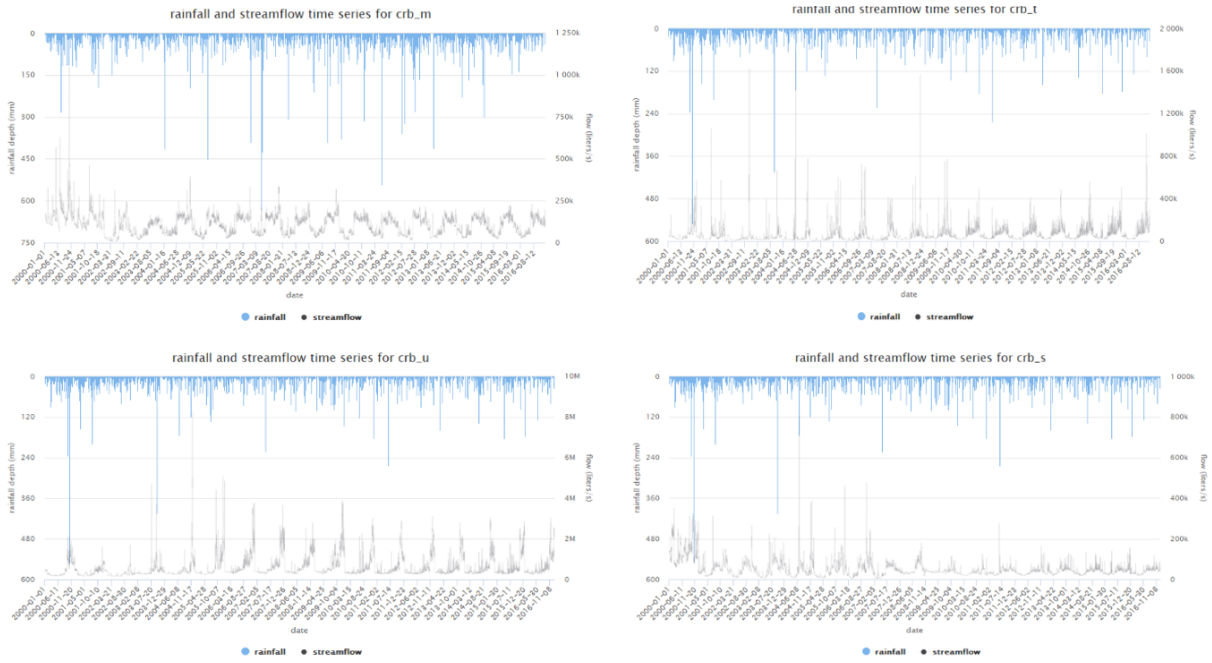


Figure 21: Subwatersheds representing the group that are not rainfall-responsive.

6.3.2 Start-end year comparison

The average predicted streamflow per season was compared (Figure 22) for the starting and ending year (2000 and 2016). Generally, most subwatersheds showed decline of water flows in dry season (18 out of 21) and otherwise during wet season (14 out of 21). Relatively, there was an average of 148% decrease in dry season and 7% increase in wet season, respectively. The highest changes for both seasons was observed in subwatersheds arb_b, crb_d, and prb_b. These subwatersheds were relatively small and with no water regulating structures.

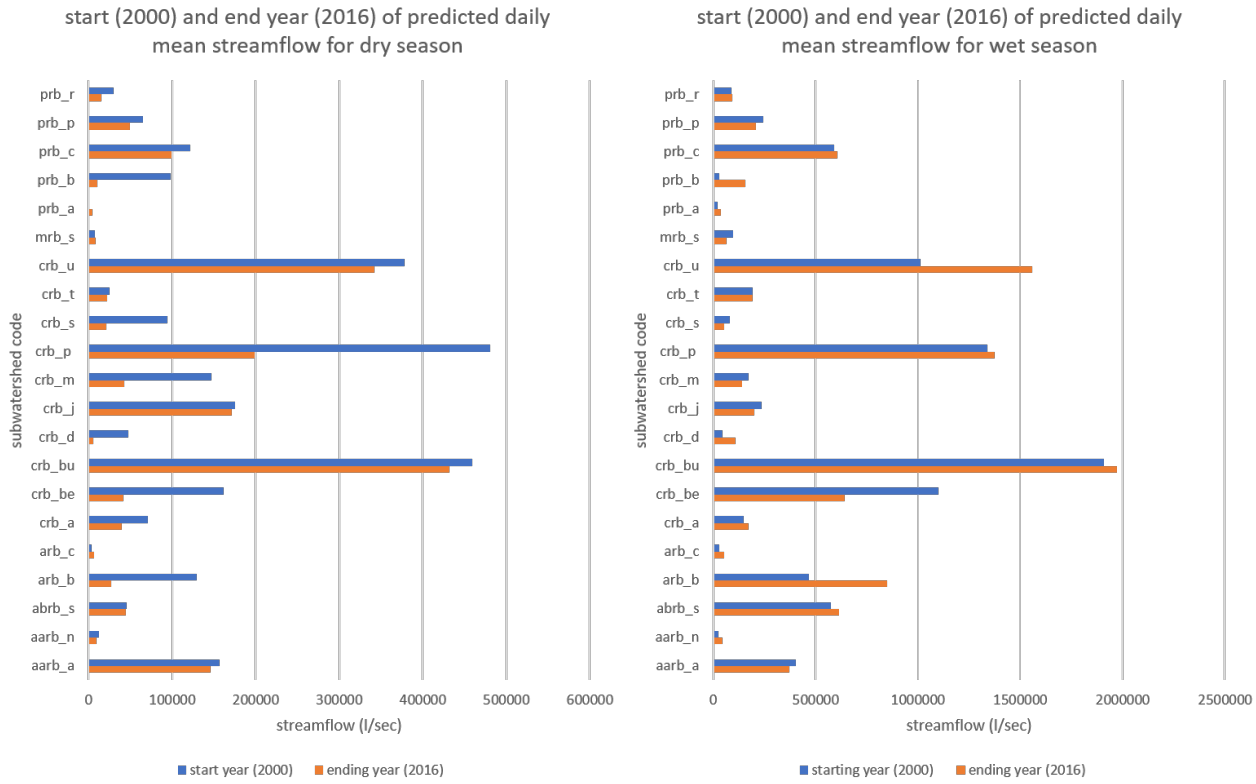


Figure 22: Dry and wet season comparison of flows in 2000 and 2016.

6.3.3 Forest loss effect

Deviation from historical flows

The two figures (Figure 23 and Figure 24) display the comparison of the seasonal predicted daily mean streamflow to its historical average (1980-1998) for subwatershed arb_b. The chosen subwatershed depicted the FLAC effect better than the other 2 subwatersheds although all showed the same effect. Note again that there were only 3 subwatersheds with historical data.

The first figure (Figure 23) shows how the seasonal flows differed without FLAC. Generally, the time series trend was more obvious from the seasonal predictions with FLAC. During dry season (upper images), there was a decreasing trend with fewer fluctuations over the years relative to the historical flow when the predictions included FLAC. Oppositely, there was an increasing trend during wet season for the FLAC predictions while extreme fluctuation was observed from the predictions without FLAC. Remember that this observation occurs also in Figure 17 where outliers appeared from models without FLAC. This subwatershed (arb_b) has a very pronounced dry season so rainy summer days are very unlikely. Moreover, there was no extreme typhoon from the area in 2004 so the high fluctuation observed in lower right graph is also unlikely.

The next figure (Figure 24) presents how the seasonal predicted daily mean streamflow deviated with the historical flows per season. The sub-graphs belong to subwatersheds with historical data namely arb_b, crb_d, and crb_m. The graphs on left are for dry season while the other side is for wet season. In general, the dry season graphs depicted a decline in streamflow while wet season graphs portrayed otherwise relative to the historical mean flow threshold. The most deviation was observed in crb_d, a large subwatershed with no regulating structures, as shown by the middle graphs. Another extreme change is shown by the upper right graph (arb_b, wet season) which started in 2003, also the same

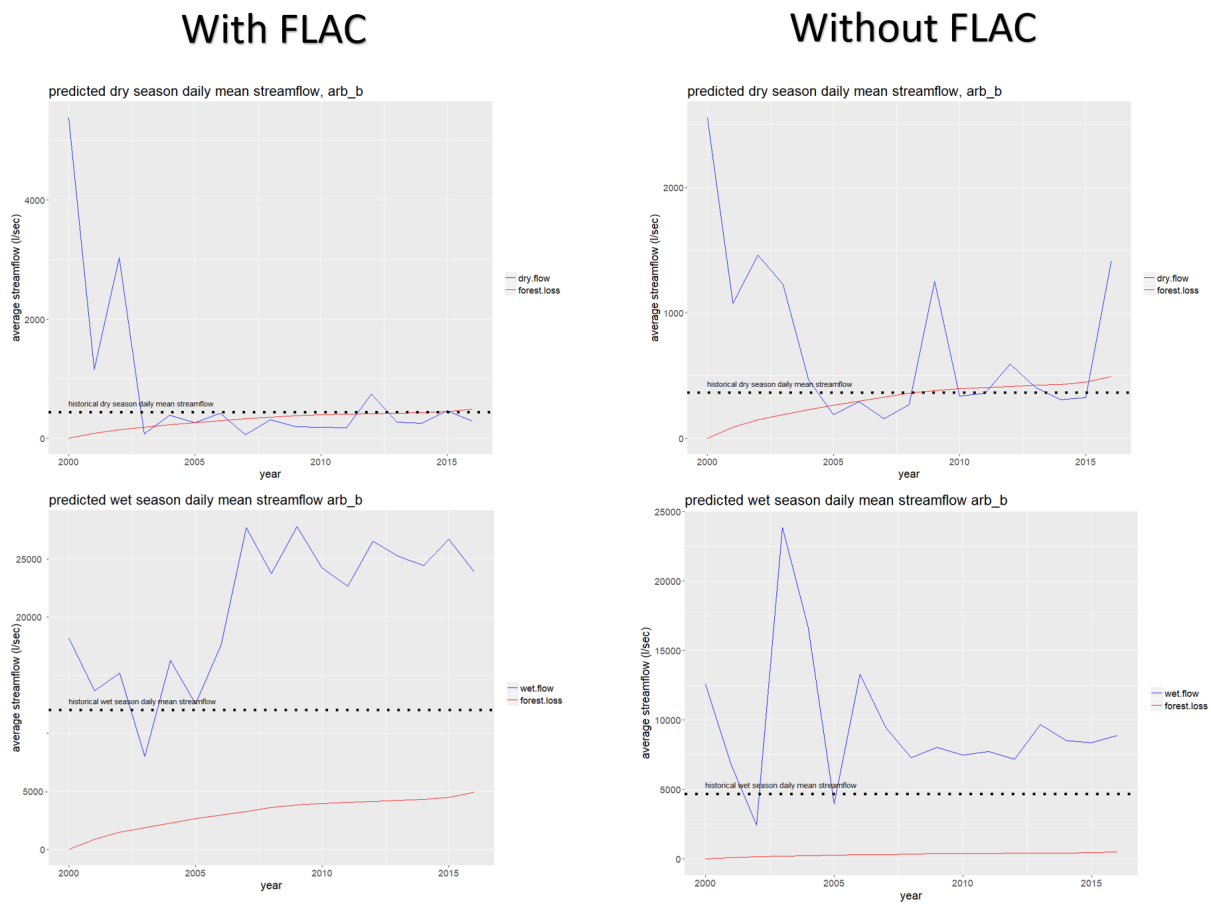


Figure 23: Predicted mean daily dry and wet streamflows (above and below images) with and without FLAC (left and right images) with the historical mean daily dry and wet streamflow threshold.

year when the subwatershed started to be below the historical dry flow threshold. This subwatershed was neither a subwatershed with regulated infrastructure nor a poor model result. As forest loss happened, only one year was above the historical dry season average flow except for subwatershed crb_m mainly because it has water regulating structures.

Forest loss and dry season streamflows

An inverse linear trend between daily average dry season streamflow and forest loss was observed for 14 out of 21 subwatersheds. Figure 25 shows the 6 subwatersheds with the strongest link or highest R^2, the rest are shown in (Annex K). Out of the 6 chosen, 4 were without water regulating structures (arb_b, crb_d, prb_c, prb_r) while 2 were with (crb_be, crb_m). Note that the dots from each sub-graphs were 16 in total depicting the full-scale prediction from 2000 to 2016. The 4 subwatersheds without regulating structures incidentally had an average of 0.76 R^2 while the 2 subwatersheds with regulating structures had 0.40 from its accuracy test. It means these 6 subwatersheds are generally good performing from the ranks of without and with regulating structures, respectively. The inverse trend from the 6 graphs interprets that there is a link between higher forest loss and drier water flows. Moreover, it indicates that even subwatersheds with regulating structures can suffer incremental water shortage during dry season.

Similarly, the 6 validation sites depicted an inverse relation between forest loss and dry streamflow as shown in Figure 26. The trend were strong in subwatershed prb_vl and abrb_vs, both have a distinct

dry season according to its climate type.

Rainfall - peak flow ratio (RPFR) and forest loss

The ratio of rainfall and peak flow is a relative measure on how peak flow is changing over time. The lower the value means there is relatively higher streamflow given a certain rainfall. The 3 most rainfall-responsive subwatersheds (aarb_n, crb_d, crb_p) showed significant relationship with forest loss with 0.52, 0.90, and 0.72 R^2, respectively. The results for the rest of all subwatersheds is shown in Annex L. The said subwatersheds happened to be unregulated subwatersheds also. The subwatershed with the highest correlation (crb_p) showed a distinct decreasing trend of the ratio. This means

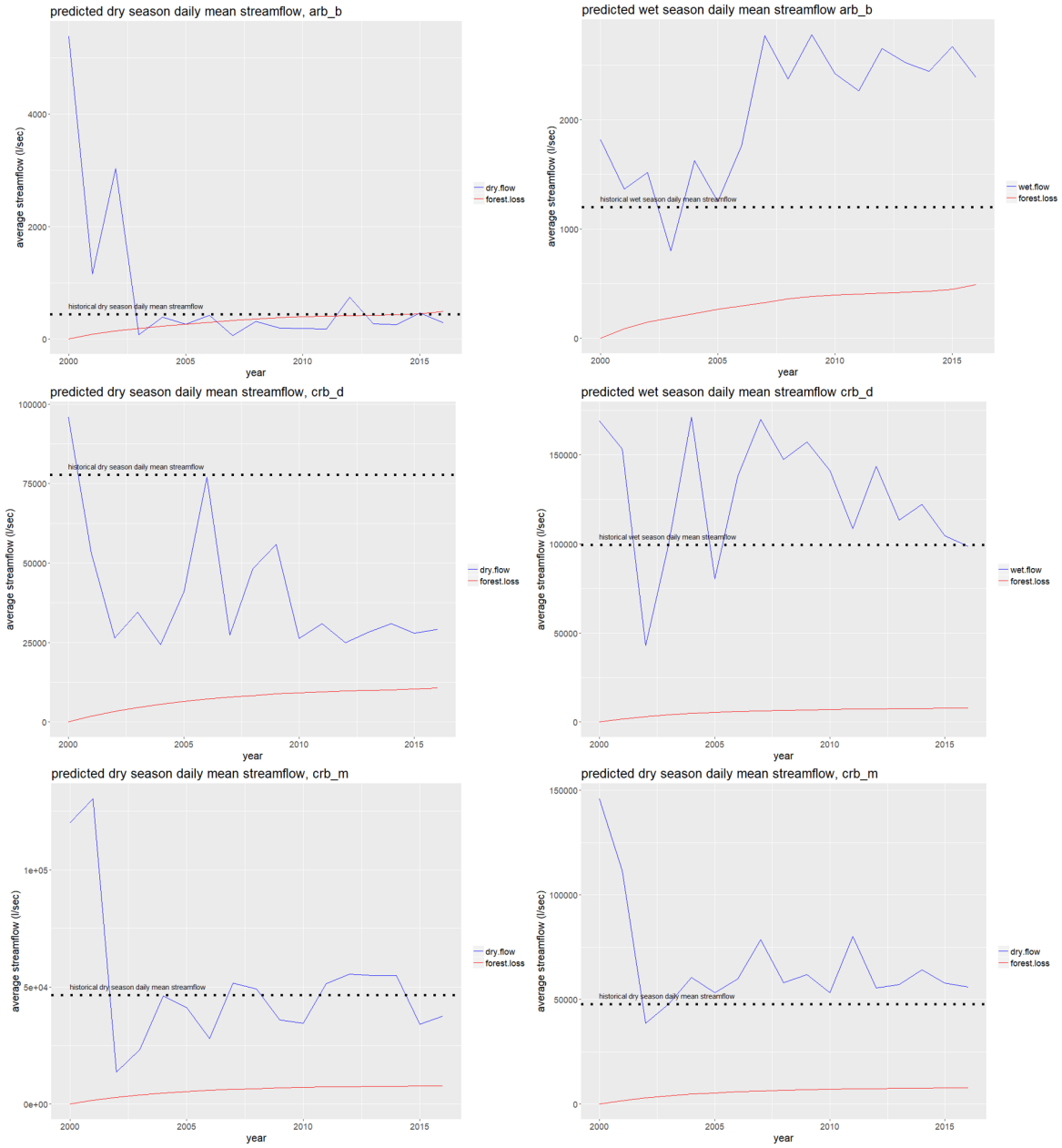


Figure 24: Predicted dry (left images) and wet (right images) mean daily streamflow compared to its historical mean for 3 subwatersheds with historical data.

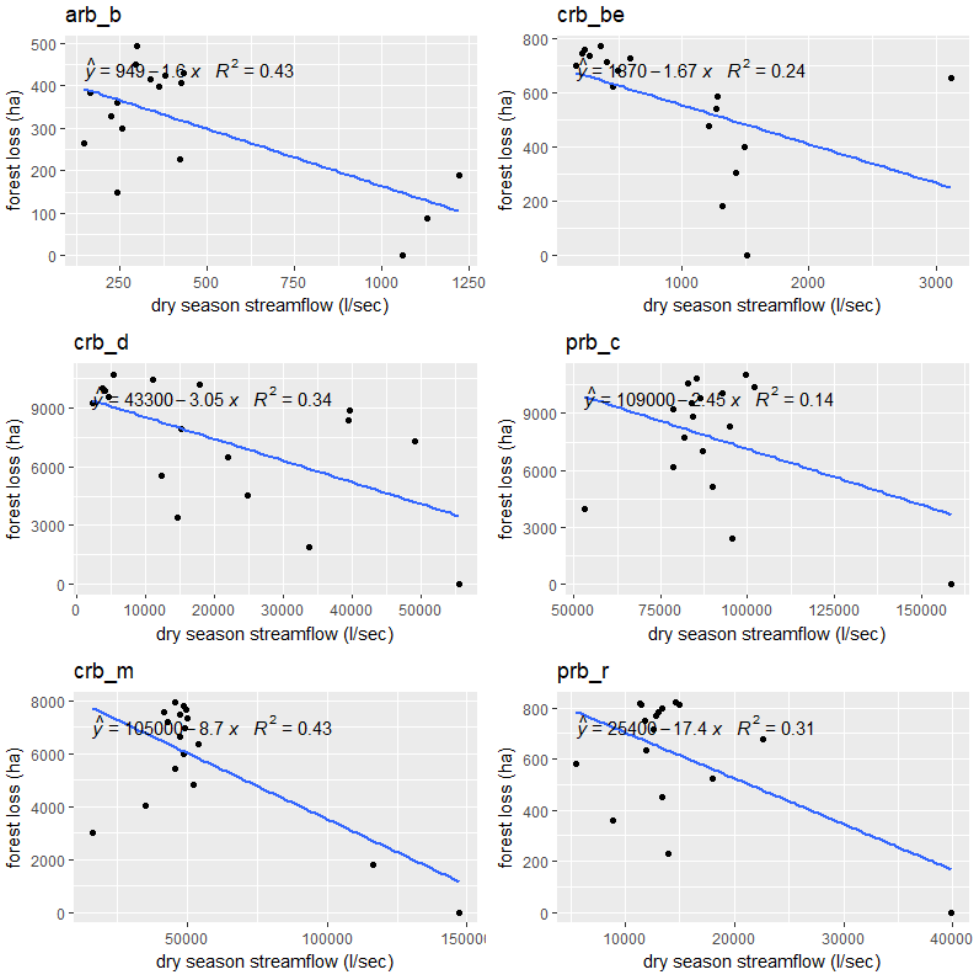


Figure 25: Mean dry flows and forest loss regression graph for 6 subwatersheds with strongest correlation.

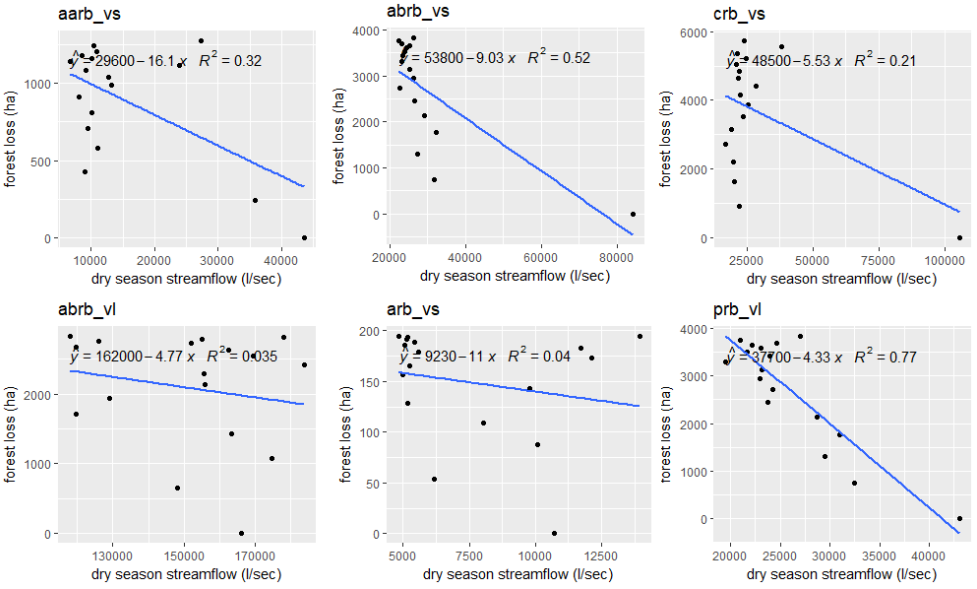


Figure 26: Mean dry flows and forest loss regression graph for the 6 validation sites.

low-high forest loss is correlated with a high-low RPFR.

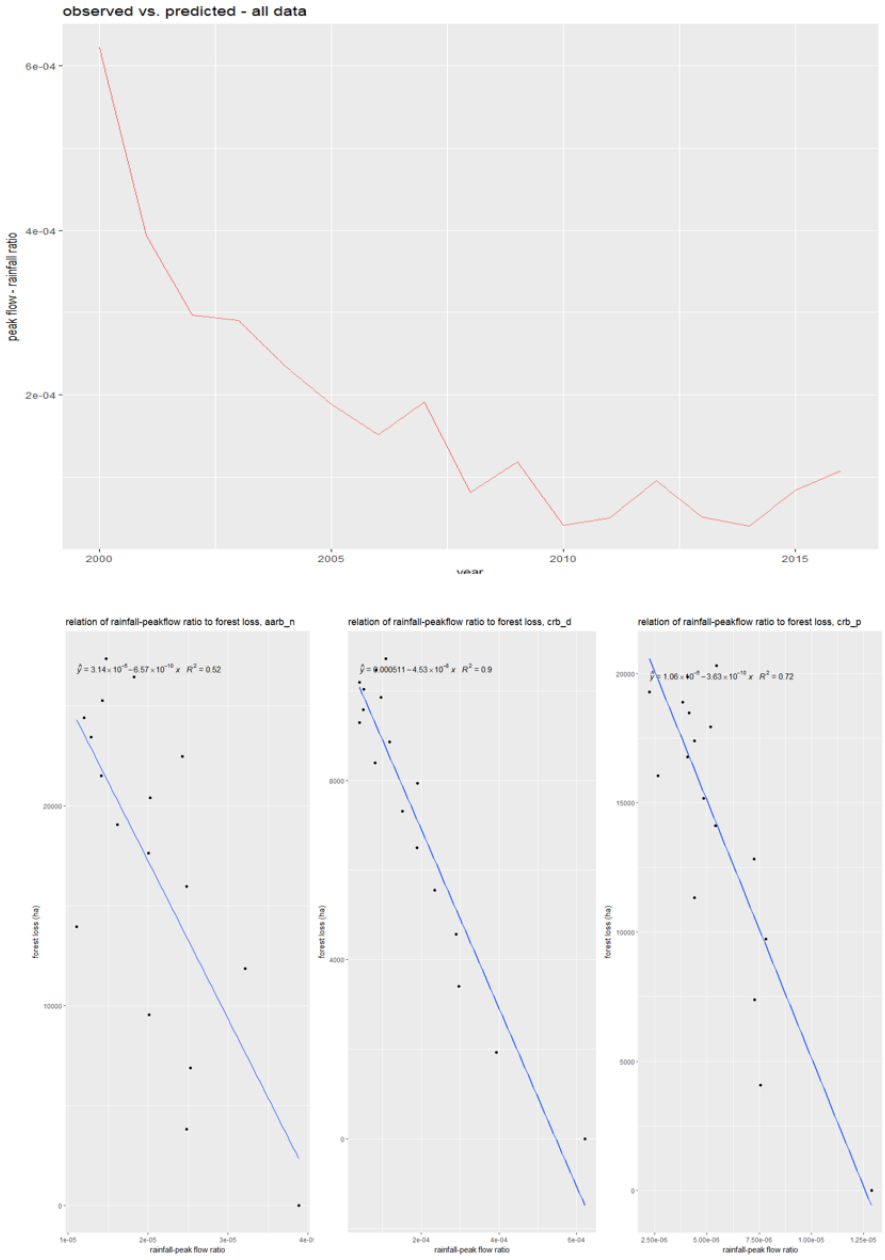


Figure 27: Rainfall-peak flow ratio and forest loss relationship.

Forest loss rate scenarios

A mean of 34% decrease in dry season daily streamflow was observed retaining the current rate of forest loss at 3%. Streamflow decreased more at 41% and 48% when forest loss rate is increased by 5% and 10%, respectively. The results had a confidence interval of +- 14%. Subwatersheds aarb_vs, arb_b, and prb_r were the only areas above 30% decrease for all 3 scenarios. These are relatively small subwatersheds. On the other hand, subwatersheds with lowest % decrease were crb_m and crb_d, both large subwatersheds with 71,897 ha and 111,512 ha remaining forests.

On the other hand, peak flows was more affected by forest loss by exhibiting higher % increase by having a mean of 88% increase with a confidence interval of +- 5%. Scenario-wise, the crb subwatersheds exhibited significant increase doubling the peak flow indicator (RPFR) when forest loss rate is increased by 10%.

Shown in Figure 28 are the scenarios for both dry and wet season, indicated by % decrease in streamflow and % increase in peak flows (RPFR). The supplemental information on forest area, subwatershed area, and forest areas from the 3 scenarios are shown in Annex M).

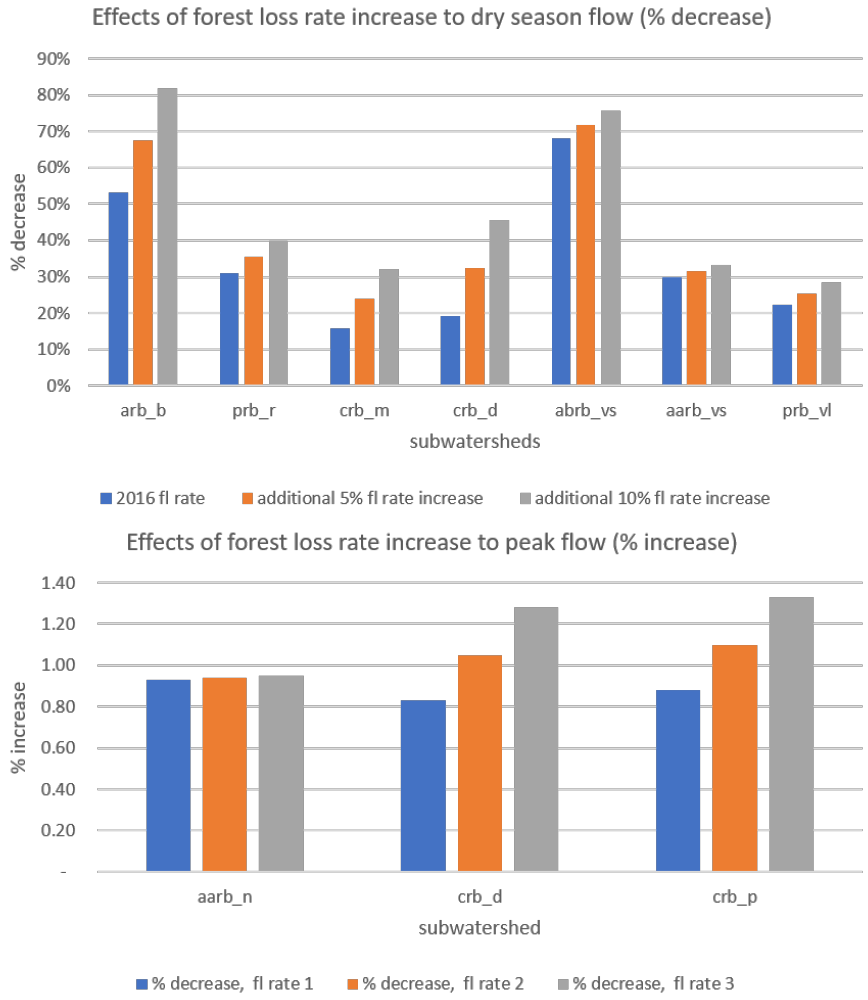


Figure 28: Forest loss rate scenarios effect to dry season flow and peak flow.

Precipitation and seasonal flows

The same regression procedure was applied to precipitation and seasonal daily mean streamflow to evaluate their relationship. It was found out that precipitation and the seasonal flows (both dry and wet) did not have any relationship at all. All regression results yielded very low correlation outcome with an average of 0.009 in R^2.

7 Discussion

7.1 Grouping-learning RF models accuracy

7.1.1 *Valuetable* grouping effect

The learn-all method attained the highest mean accuracy result for both R^2 and NSE among the grouping-training methods. This finding indicates that supplementing the RF algorithm with all information coming from the 21 subwatersheds can be the best way of model training. However, the basin-level method did not perform beyond the learn-all method. In fact, only one basin was below the R^2 threshold of > 0.60. This means that as long as there is sufficient grouping of information even at basin scale, the RF algorithm can learn and predict fairly well. This finding was not found for two relatively poor performing grouping-learning methods: the subwatershed-level and the small-size subwatershed grouping. The main reason for the former is because no "outside information" are included from the model (since no *valuetable* grouping happened), as such, static covariates namely elevation, slope, and soil are nullified. The insignificance of the static covariates were reflected from the VIM graphs by having 0 importance values. Another reason why the subwatershed-level method performed poorly is that the subwatersheds are located in different parts of a larger basin exhibiting different water flow properties relative to its location. Given that the static covariates are out of the picture, it become insignificant during the model training process. Hence, the model lacks information from the following characteristics: (1) a more sloping subwatershed has higher surface runoff because of gravity (Engman, 1986), (2) upstream subwatersheds have the least streamflow by having no influx or incoming water to it (SWAT Documentation, 2012), (3) subwatershed in high altitudes have more rain (Winiger et al., 2005), (4) clayish soils are less penetrable by water (USDA, 1986). The second poor performer among the grouping-learning method was the small subwatershed grouping. In essence, smaller subwatersheds are just parts of larger basins and these larger entities have more defined hydrology dynamics (Jones and Grant, 1996). Moreover, the smaller subwatersheds belong to 3 different climate types while larger ones are confined to 2 types. That can amplify climate variability within the grouping method for smaller subwatersheds. The size classes could have been 5 classes to have more distinction among subwatershed sizes, however, 21 subwatersheds are insufficient to have enough number per size class.

7.1.2 Number of observations effect

Higher number of observation was not always equivalent to higher accuracy result. This was observed from the basin and subwatershed grouping-learning methods where the basin and subwatershed with the most number of training data just got 0.58 in R^2, on the average. There seems to be a tipping point where the RF algorithm fully learns the area of interest regardless of the number of observation. Again, the degree of variability among the subwatersheds and even basins cause this. For example, the RF algorithm can learn better from a subwatershed with very distinct wet and dry season rather than a subwatershed with twice the number of observations but is more heterogeneous physically and climatically. An example of this are subwatersheds arb_b and prb_r attaining the highest accuracy result and both belong to climate type 1 (distinct wet and dry season). But when the RF algorithm starts to distinguish different characteristics of basins and subwatersheds, the number of observations make more sense because of the grouping process. That is why the learn-all method performed well because it has the most "outside information". In other words, combining information from various subwatersheds is more sensible than having more training data for one subwatershed or basin.

7.1.3 Water regulating structures effect

Model performance was affected by the presence of water regulating structures within subwatersheds. Logically, those artificially regulated flows can behave differently in regression models (Shortridge et al., 2016). At subwatershed-level, there were 11 out of 21 that performed below the significant R^2 threshold. Most of the poor performing subwatersheds (9 out of 11) belong to basins crb and prb where most water regulating structures are located (Annex I). Most of it are dams for irrigation and flood control structures since within the two basins are top producing rice provinces in the country (PSA, undated) but also the most flooded (DPWH, undated). At basin-level, information from subwatersheds are mixed including those with and without regulating structures. This can also be the reason why those information-rich basins like crb and prb, which combine for almost 70% of the total number of observations, did not attain the highest accuracy results. These structures, when operating in smaller subwatersheds, have more effect on streamflow rate than that of large subwatersheds because practically of the size difference. For instance, dams are mostly located in the main stream of small subwatersheds while it is just in tributaries for larger subwatersheds. Most of the regulating structures are constructed in small subwatersheds which can be the reason why size-based grouping yielded lower result for small subwatersheds. Moreover, it can be observed also from the box plots of the small subwatersheds (Figure 19) that there was more balance on wet and dry flows compared to unregulated subwatersheds with more rainy day flows (which is natural in Philippine climate). As said, this can be attributed to the release of irrigation water during dry cropping season. Moreover, farmers tend to have 3 cropping seasons instead of the traditional twice a year leading to more irrigation water use. The RF algorithm could have learned this if a covariate distinguishing with and without regulating structures was included. Otherwise, a separate grouping-learning method for with and without these structures could be made. Other information specific to the operationalization of the structures could also be supplemented as this is highly variable given a 16-year span. For example, irrigation systems in the Philippines are always being repaired either because of typhoon damage or substandard construction. But still, despite mixing the information, the RF can manage to learn artificial water regulation in an extent. The number of subwatershed with and without regulating structures were 11 and 10, respectively. Because of that balance, the learn-all model have more capability to distinguish it than other grouping-learning methods.

Given the said findings, it can be summarized that combining each subwatershed information, even with different characteristics, and taking information from both with and without regulating structures can still yield a significant accuracy result but not the highest.

7.2 Predicted streamflow

7.2.1 Hydrograph assessment

Hydrographs or streamflow-rainfall graphs were used to assess full-scale (2000-2016) predicted streamflow. A watershed can be assessed by being "rainfall-responsive" or a discharge behavior in accordance to a rainfall event using hydrographs (Akhtar et al. 2009). There were two distinction among the 21 subwatersheds as per hydrographs: those rainfall-responsive and not. Assessing the former is visually obvious since rainfall-streamflow has a linear relationship meaning high rain is equals to high streamflow and vice versa. Additionally, rainfall-responsiveness can be attributed to the accuracy measures as it was found out that rainfall-responsive subwatersheds were also those above 0.60 R^2. However, for the latter, assessment can be more complicated. One reason why a subwatershed is not rainfall-responsive is because of water flow alteration. A similar observation is reported by the study by Shortridge et al., (2016) where the model did not capture the rainfall responsiveness of

the discharge. They concluded non-linearity between rainfall and discharge because of dampening effect of wetlands and floodplains. As discussed from the previous paragraph, the presence of water regulating structures alter water flow. This behavior was observed in the hydrgoraphs visually and it was evident from subwatersheds crb_t and crb_m where dam operationalization was traced from the rainfall-streamflow patterns. However, despite capturing dam water regulation in an extent, a stringent attribution to it still needs in-depth assessment with inclusion of operationalization dates and water release mechanisms. There are protocols followed by these structures i.e. water release at certain dam water height. Nevertheless, it can be concluded that hydrographs can give clues which subwatersheds have water regulating structures.

The predicted streamflows from rainfall-responsive subwatersheds can be used for quantifying the water services. Since these are dynamic outputs (daily), it can be integrated with other modeling techniques. Specifically, results of this methodology can be integrated with process-based models and individual or non-aggregated models (Cellular Automata and Agent-based Models). For the former, predicted/forecasted daily streamflow can be used for flood modeling. For the latter, social component like ecosystem stakeholders (i.e. government offices, farmers, local government), can be included in the streamflow modeling outcome. This is relevant at smaller scales i.e. municipal-level where the focus of planning and management takes place. The local government units are mandated to enact local plans which include scenario building for the future (25 years at most). Integrating ecosystem services in that picture can make their plans more oriented towards the environment in a long-term basis.

7.2.2 Peak flows assessment

Peak flows from extreme rainy days or streamflows with relatively high values were under-predicted. In fact, this finding was common to all models of different grouping-learning methods. One logical reason of this shortcoming is that the extreme flows deviate from the mean flow (for the whole training data) by far and these flows occur seldom and irregular. The study area is having around 3.7 typhoons per year (Silent Garden, undated) and by being seldom, is treated by the model as outliers. Another reason is the lag-time effect or the time difference between peak rainfall and peak flow which can occur and be recorded in two adjacent days especially if the peak rainfall occurs during nights (Granato, G. E., 2012). Mirroring this case are post-typhoon days where subwatersheds still have high streamflow rate even without rain anymore (Abon et al.m 2011) or essentially a "no rainfall-high streamflow" scenario. Those events essentially contradicts a "high rainfall-high streamflow" combination resulting into an "averaged prediction" from the final prediction of RF trees with these cases. Midway this study, this limitation was addressed when the maximum daily rainfall covariate was added, the peak flows prediction improved even just using one test subwatershed. This proves that the predicted flows are really sensitive to rainfall and can overcome the prior limitation of under-predicting extreme flows.

Despite the under-prediction of peak flows, a distinct trend was observed from the unregulated subwatersheds from the RPFR results: peak flows were increasing from 2000 to 2016 during extreme rainy days. This is an important result since the RPFR is an indication that more streamflow can occur over the same rainfall magnitude over time. This can be tagged to the flood regulation service of forests, however, that service cannot be quantified (in volume) based on this trend alone. Nevertheless, it can be modeled using the methodology of this thesis if sub-daily weather and streamflow inputs are to be integrated.

To discuss shortly about the importance of peak flows, around 80% of the area are considered flood plains (Philippines DRRM, undated). Peak flows are important in assessing flood regulation service

of forests during these events (Hein et al., 2006). Existing studies concluded that a more forested watershed can reduce impact of flood as indicated by lag time and streamflow volume (Calder and Aylward, 2006; Aquino et al., 2014; Rawlins et al., 2016). The main indicators for the said forest service are: delay hours of flood water, influx decrease, inundation area decrease; and flood height decrease.

7.2.3 Average monthly seasonal flows assessment

More smoothing and fitting with the observed data was noticed from the predicted streamflow line graphs when averaged per month. Moreover, the comparison between predicted and observed streamflow was depicted better in a monthly basis rather than daily. The smoothing is mainly because of the averaging effect of daily streamflows that included nullification of those from extreme rainy days. The monthly averaging also revealed that there was more over-prediction during wet season (14%) than in dry season (6%) from the simulated streamflows. The main reason for this is that there are more wet than dry days in Philippine climate. Moreover, watersheds in mountainous tropics can have its own micro-climate (Wang, et al., 2010) which defies the law of the general weather pattern by raining more. The RF algorithm takes the final vote of each prediction tree and averages it. Given there are more rainy days from the observed data, the RF algorithm has technically more "rainy days learning" which somehow give bias leading to streamflow over-prediction. Aside from that, the RF algorithm also captures two forms of those "pseudo rainy days": (1) summer/no rain days where the regulating structure release water and (2) post typhoon days that give "no rainfall-high streamflow" as discussed from the last paragraph. Moreover, the training data is confined until 2010 only because of data availability constraints. Climate change effects on weather is felt more in the past decade (Cruz et al., 2017) with aggravated rainy season and drier dry season including irregular typhoon and even monsoon patterns. Despite the slight over-prediction, the monthly average seasonal flows are helpful in water yield quantification indicated by water users or service area. For example, monthly streamflow can be sufficient for quantifying water requirement per cropping season of irrigated crop lands. It can be used also in setting water usability threshold like Dependable Flow (DF) or the threshold for "usable" water volume. Usually this measure is standard to tropical subwatersheds belonging to same climate type and is set by water regulatory institutions. In the Philippines, common DF should be around $>$ 80% (Apayao-Abulug Master Plan, 2015).

7.2.4 Validation outcome

The validation results for 6 subwatersheds depicted both seasonality and rainfall-responsiveness. This is one key result of the study indicating that the RF model can be applicable to subwatersheds without observed data. This further means that the RF model can be used to upscale to other places preferably to those within the Luzon island. Note that the 6 subwatersheds are pre-selected to have no regulating structures. It would be interesting also to validate for those with regulating structures in the future given the abovementioned suggestion to include a covariate or a separate grouping-learning method accounting these structures. The validation results can be further compared with other studies and master plan, however, time is scarce for that attempt.

The outcome of the validation can depict homogeneity in the study area. Logically, a more homogeneous area can be learned better by the RF model than a more heterogeneous area. Imagine a heterogeneous area with hundreds of covariates resulting to tens of thousands of RF trees which is possible for this study but impractical. Limiting the covariates for this study also over-fitting from using highly correlated covariates (Kroll et al., 2004) which are common to hydrological parameters. Moreover, it should be noted that 7 dominant land cover classifications were used for this study (forests, perennial crops, annual crops, built-up area, grass land, water, and bare land). The static

covariates like soil, slope, and elevation were reclassified into 5 classes. It was made to save processing time since the static covariates have continuous data values. In terms of climate, the study area fell into 3 different types of climate. The said aggregations made for this study can be practical based on the "analysis scale". It defines how the aggregation can be and given the study area is at basin scale >15,000 km^2, aggregation of spatial inputs are more possible. Despite the aggregation, however, the inputs used for this study did not sacrifice spatial and temporal resolution.

7.3 The forest loss effect

There were three main effects of FLAC to the RF models and its predicted streamflow: (1) the linear trend between forest loss and seasonal flows depicted by two indicators: *(a) the PDPs* and *(b) the correlation of forest loss to dry flows and peak flow-rainfall ratio*; (2) regulation of over-predicted streamflow indicated by the *(a) deviations from historical flows* and *(b) actual/observed streamflow*; and (3) reduction of RF model unexplained variance.

7.3.1 The linear trend of predicted streamflow and forest loss

Based on PDP

Almost all PDPs of subwatersheds (20 out of 21) exhibited low-high (negative linear) and high-high/low-low (positive linear) link between streamflow and FLAC, respectively. Remember that linear PDP graphs whether inverted or not both exhibit a strong effect with the prediction of streamflow. A plausible explanation is that forest loss covariate being an absolute cumulative value is reflected four times (since there are 4 FLACs) and in fact, that trend is exclusive only to FLAC among 58 covariates. Meaning to say, there are high chances that $>= 1$ of FLAC is/are always included to the RF trees. The high number of *mtry* RF parameter assured that this trend is always produced in every RF tree. As such, high chances also that the distinct trend of FLAC values would coincide the distinct trend of the dependent variable streamflow. It would be interesting to assess also the % of trees where FLAC is included and relate it to the prediction using test data, however, there is no easy way to extract a tree-level information from the random forest function.

Based on correlation

Three key findings were linked to assess the effect of forest loss to the predicted seasonal streamflow. First, there was a 148% decrease and 7% increase, on the average for all subwatersheds, in dry and wet predicted daily mean streamflow, respectively. Second, 14 subwatersheds exhibited an inverse relationship between mean dry predicted streamflows and forest loss. A 16-year stretch could be insufficient to perform a decent regression analysis giving 16 points to correlate. Nevertheless, 6 among 14 subwatersheds showed an inverse and decent regression result with an average of 0.32 R^2 between forest loss and dry flows. Incidentally, these subwatersheds were good performing models, accuracy-wise having an average of 0.76 R^2. Third, the RPFR of all unregulated subwatersheds signified an inverse correlation with forest loss. The highest subwatershed attained 0.90 R^2. Aside from the linearity, the RF model proved that it can be rainfall-responsive during extreme rain events. The said findings can attest that the RF algorithm can learn that increase-decrease of streamflow can be a function of forest loss. Specifically, it can have **3 interpretations**:

First, more forest loss can lead to lesser water in dry season. Theoretically, that behavior can be perceived as the water regulating service of forests that assures water flows in summer days. This

is caused by the base flow (also called drought flow and groundwater recession flow) or the form of water contributing to the streamflow during dry weather. In a study by USGS in 2017, it was concluded that a more forested watershed can have more base flows especially during dry season with no rains compared to a pastural-agricultural watershed (Rodriguez and Santiago, 2017). Moreover, the trees within the forests of the study area are mostly indigenous or native trees (DENR-FMB, 2016) which are found to have no water seeping effect that dries a watershed unlike exotic or introduced trees (Zhou et al., 2002). Drought was also experienced within the service area in 2010 and 2016 (Taguinod, 2010 and The Northern Forum, 2016). Deforestation is aggravating the effects of climate change (Staal et al., 2016) and prolonged drought is one of its outcomes. The Philippine government was aware of that so total logging ban was enacted several times during the span of 2000 to 2016. The government also irrigated 300,000 ha of farmlands during the 2006-2010 span (BSWM, undated). Moreover, there was a massive national reforestation of indigenous trees happened in 2012 called the National Greening Program (NGP) by the Philippine government which targeted 1.5 million hectares of reforestation. Assuming a successful survival of the transplanted seedlings, these are still poles ($<$8cm diameter) until 2016 and may not have a major hydrological impact, thus, its effect may not yet captured from this thesis' results. However, it would be interesting to capture its impact 5-10 years from now using this thesis' methodology by integrating forest gain pixels yearly.

Second, more forest loss can lead to higher surface run-off in extreme rainy days based on peak flows. Forests provide an important service during extreme rainy days i.e. typhoon and monsoon season. The three main functions of forests during these events are increasing soil infiltration rates, reducing run-off, and preventing soil erosion (Smithers et al., 2016). It should be noted that the forest pixels used in this study were net forest loss meaning, forest gains were deducted already. Whatever land cover preceded after forest loss, it seems forests can provide the least surface run-off when raining extremely. This is justified in literature from the so-called "rainfall-runoff" ratio (USDA, 1989) which is very similar to Curve Number in terms the values per land cover. It is a coefficient how run-off is affected by the vegetation cover and soil. Moreover, it is also similar to the RPFR in terms of values but inversely. Various reports in news articles from https://reliefweb.int/organization/govt-philippines affirmed this as stakeholders are blaming forest loss to hastened flooding especially in basin crb. To remedy this, the government spent billions of pesos in the past decade for flood control structures within Luzon (DBM, undated). These structures can be the reason why the relative % increase of predicted wet season streamflow is just 7% between 2000 and 2016. Moreover, the Philippine government published high resolution rainfall projection in 20, 50, and 100 years (PAGASA, 2016) out of climate change effects. One of its highlights was increase in rainfall projection in wet season. Assuming that deforestation would continue, its impact could worsen the effects of climate change in the study area especially on flooding. It would be fit for this thesis' methodology to forecast those climate change scenarios by changing the weather covariates. Moreover, deforestation rate scenarios could be created to have support decisions on flood infrastructure construction and rehabilitation by the government.

Third, water regulating structures can have functional limitation. Interestingly, decrease in water flows during dry season for those subwatersheds with water regulating structures (regulated subwatersheds) was evident for 6 subwatersheds. It indicates that water shortage can still be possible even for subwatersheds with dams and irrigation canals. On the other hand, during wet season, 4 subwatersheds with water regulating structures experienced water flow increase. This means that the water storage capacity of these infrastructures especially in rainy season can still be insufficient. One main reason for this is sedimentation which is a function of forest loss and is amplified when forests turn to bare lands (David, 1988). Desiltation efforts can be costly especially for dams and irrigation

canals. Moreover, dams tend to release water during successive rainy days when storage hits its limit.This scheme can be irregular and varies depending on the protocols of the dam managers. It further validates that indeed the behavior of subwatersheds with regulating structures are way more complex than those without.

The forest loss effect to seasonal flows cannot be mis-attributed to rainfall trend. Remember that weather information was also regressed with the seasonal flows since changes in rainfall pattern yearly can reflect on yearly streamflow. It appeared that rainfall trend has no effect at all as shown by the very low correlation with the dry flows. Moreover, typhoon records supported this as typhoons appeared irregularly and not sequential from 2000 to 2016 (PAGASA, 2016 and de la Cruz, 2016).

Despite the "linear" findings between forest loss and the streamflow indicators (seasonal flow, RPFR, and PDP), the significance level of the correlations were not very high i.e. 0.90 R^2 and above. This can be attributed to two things: **First**, the preceded land cover/use after forest conversion and its land management practice. Upland farming in the Philippines is very common and it is a "magnet" for communities to migrate in forest areas and do forest conversion into agricultural lands. Croplands (especially perennials) can also regulate water but relatively lower than forests (Rawlins et al., 2017). The government was promoting Agroforestry or combining forest trees and agricultural crops in upland areas since early 90s which is perceived to be more sustainable and anti-soil erosion (Baguinon et al., 2007). Forests turn into bare land in the Philippines by 4 common causes: (a) clear-cutting out of logging, (b) slash and burn, (c) mining, and (d) forest fire. Often, those logged out areas tend to regenerate after decades. However, those bare lands out of the other 3 causes result into grass lands and wooded grasslands. As early as 1998, the government started investing for bio-interventions like vetiver grass, a known grass for soil erosion control (Manarang et al., 2000), to cover erodible bare lands. This grass is being planted both on steep barren slopes and river banks and by being a "grass", reflects and gets classified as grass lands in land cover. But, it has more hydrologic function than a regular grass land because of its soil holding capacity and run-off regulation. Similarly, there are coconut fiber mats (also known as "erosion blankets") used for the same purpose as vetiver grass that hold soil. These projects were implemented within the scope of the study area (BSWM, undated; DENR, undated; Rawlins et al., 2017). The report by Rawlins et al., in 2017 highlighted the immediate erosion regulating effect by coco mats than reforestation. Again, these areas can be classified as bare land, however, function differently as far as hydrology is concern. As abovementioned, including the land conversion to this methodology would make it more definitive. Moreover, including the land management practice given sufficing data could make the methodology more stringent. **Second**, the forest pixels used for this methodology were not perfect. According to its source (Hansen et al., 2013), there was an 83% and 87% of producer and user accuracy, respectively, on forest loss for tropics. As such, there is still room for misclassification. Commonly, forests are mistaken as perennial crops/plantations and wooded grassland, vice versa (NAMRIA, undated). Forest gains, on the other hand, depicted lower accuracy results by just having 82% and 48% on producer and user accuracy, respectively. Now that there are more accessible high spatial and temporal resolution optical images like Landsat and Sentinel (2A and 2B), land cover classification can be more doable yearly. Moreover, fusion with radar images can deal with the cloud problems of optical images (Joshi et al., 2016). Moreover, the Philippines has a nationwide LiDAR data acquired via DREAM-LIDAR program of the government. It created very high resolution agricultural maps (https://parmap.dream.upd.edu.ph/). These maps can be integrated in creating future land cover time series in the future using this methodology. Moreover, using time series algorithm like Breaks For Additive Season and Trend (BFAST), future forest loss time series classification could be automated and integrated to this methodology.

The linearity between forest loss to dry season streamflow and peak flow were used to forecast an increase in forest loss rate. Streamflow in dry season decreased while peak flows doubled in values when 10% of forest loss rate was added to the current rate. It was found out also that smaller subwatersheds were more sensitive to forest loss than larger ones for dry season scenarios. This can be attributed to the impact of remaining forests to smaller subwatersheds where rivers can dry faster than those larger ones especially if the area has relatively lower forest %. The finding of Ogden et al., (2013) supports this as they found out that runoff from the forested subwatershed receded more slowly than from the a mixed land use and pasture subwatersheds. The irrigation priorities in the Philippines are those prone to this phenomenon (NIA, 2011) where river-drying technically known as baseflow drying are happening often in smaller subwatersheds. As such, the small-scale irrigation system was promoted by the government known as SSIP where the focus areas were those drying-prone rivers in small subwatersheds. On the other hand, the size of the subwatershed has an opposite impact on peak flows increase when forest loss increase. Larger-forested subwatersheds can be more sensitive to deforestation during extreme rainy days with increasing peak flow indicator (RPFR) or simply more water volume can be discharged given the same rainfall amount. It makes sense since in a larger subwatershed, more influx are collated to the main stream. Logically, there is more gravity pull and velocity also in larger subwatersheds. Comparatively, peak flows are more sensitive to forest loss than dry flows by having higher magnitude of change. It can be attributed also to the function of the soil where its water storage function is lost faster in extreme rainy days due to saturation given the same forest loss scenario.

7.3.2 Regulating streamflow over-prediction

Deviation to observed/actual streamflow

The effect of FLAC was reflected from the comparison of the predicted and observed daily streamflow per month. Remember that without FLAC, streamflow in wet months was over-predicted in 5 basins by 14% on the average. This was also reflected from the accuracy measures having a relative 17% decrease in R^2 and NSE when FLACs are removed. It should be given emphasis again that the 4 FLACs have the only distinct cumulative trend among the 58 covariates. These findings justify the regulating effect of forests during wet season minimizing over-prediction. It was discussed in section 6.2.3 that the RF model can have more "rainy days learning". Without FLAC, it can be more aggravated by losing a distinct trend that coincide with the streamflow trend. Adding more associated covariate like ground water and evapotranspiration variables could be more sound in the future. It could strengthen the conversion learning of the RF model by not being limited to 3 associated covariates to forest loss.

Deviation to historical mean

Similarly, the deviation from the historical wet mean flows observed in Figure 24 justified that without FLAC, there can be extreme over-prediction in wet season. The changes are considered extreme fluctuation from the historical mean and can therefore be considered outliers. A change of around 70,000 l/sec can be too much considering that the standard deviation of the historical flows is just 12,000 l/sec. Typhoons cannot cause this since no typhoon record is found during 2003 where those extreme flows occur. This observation was also evident from the daily predictions with and without FLAC using the held-out data where extreme flows appeared. Again, the impact of FLAC trend to the predictions is reflected on this finding. It can be stated that there can be more outlier-looking flows in wet seasonal flow predictions without FLAC.

7.3.3 Prediction performance effect

The RF model accuracy can decrease by 18% at basin-level when FLACs are permuted. In other words, the unexplained variance can increase by 18%. The accuracy is measured by computing the accuracy difference between two RF trees (original and permuted) and the averages it among the RF trees. Shuffling a variable with cumulative trend like forest loss can lose the linear relationship between the seasonal flows. At subwatershed-level, there were cases that FLAC tend to outperform some weather covariates. It can be attributed to again to the distinct trend of FLAC since no *valuetable* grouping happened at subwatershed-level and if this distinct trend are permuted, can worsen the models. However, this finding could be misleading in the sense that subwatershed-level can be better models as per VIM. It should be noted that the VIM of FLAC can be greater than weather covariates only in cases where FLAC have a cumulative trend and weather covariates have lesser number of observations which happened at subwatershed-level grouping-learning method.

8 Conclusions and Recommendations

Conclusions

Random forest algorithm is capable of predicting streamflow better when aggregation of *valuetable* information is applied. The learn-all model, being relatively the best model (with 0.69 R^2), outperformed other grouping-learning methods by an average of 0.2 in R^2 and 0.3 in NSE. Moreover, the same accuracy measures revealed that large watersheds are better predicted than relatively smaller ones when grouping-learning pertains to subwatershed size.

The streamflow prediction models can distinguish which subwatersheds have regulated and unregulated water flows. Unregulated subwatersheds or those with no water regulating structures like dams, flood control structures, and irrigation canals performed well according to accuracy indicators (average R^2 of 0.66). On the other hand, regulated subwatersheds reflected lower accuracy result (average R^2 of 0.23). The presence of water regulating structures are reflected also by the hydrographs. Those unregulated subwatersheds show rainfall responsiveness. Mixed information coming from both regulated and unregulated subwatersheds can still yield a significant accuracy result but not the highest. This is proven by basins crb and prb having 70% of the total training data but just yielding 0.63 R^2, on the average.

There were two error sources of the predicted streamflow common to all grouping-learning models. One was the inability to predict peak flows of extreme rainy days mainly because extreme peak flows deviate from the mean flows by far. These flows are irregular and seldom (around 4/year) therefore treated as outliers. The second error source came from over-prediction of models by around 9% especially during wet season (14%) observed from the daily average streamflow per month at basin models. There were more "rainy days learning" by RF model because of the natural Philippine climate. However, there were "pseudo rainy days" or "no rainfall-high streamflow" days learned by RF models which gave bias to rainy days. Three reasons are causing this: (1) lag time covering two days, (2) post-typhoon days, and (3) release of water by the regulating structures during days without rain.

Forest loss and its associated covariates were valuable explanatory variables contributing almost 1/4 and within the top 13 among 58 covariates on permutation importance value. Moreover, FLAC has the tendency to decrease RF model accuracy by 18% when permuted, on the average. More importantly, the RF model was capable of learning forest loss effect to seasonal streamflow based on the evidences below:

1. There is a 148% decrease and 7% increase in dry and wet predicted daily mean streamflow for all subwatersheds, on the average;
2. There are linear trends between forest loss and streamflows depicted by three indicators: (1) the PDPs of 20 out of 21 subwatersheds, (2) the 0.36 correlation between forest loss to dry season streamflow on 6 high-accuracy subwatershed models, (3) 0.72 R^2 correlation to rainfall-peak flow ratio (RPFR); and
3. Regulation of wet season over-prediction by controlling outliers and reducing over-prediction from daily streamflow average by 4%.

Using the linearity of forest loss to dry streamflow and peak flow (RPFR), forecasting scenarios revealed that streamflow during dry months decreased by an average of 48% and peak flows increased

by 114%. This is when an additional 10% is added to the current forest loss rate for unregulated subwatersheds. Moreover, smaller subwatersheds were more sensitive to forest loss while higher changes were observed in larger subwatersheds during wet season. This was based on % dry streamflow decrease and % peak flow increase (RPFR) indicators, respectively.

The predictions for the 6 validation subwatersheds both depicted seasonality and rainfall-responsiveness. This indicates that the RF model can be applicable to subwatersheds without observed data. This further means that the RF model can be used to upscale to other places preferably to those within the Luzon island.

The net forest loss of the study area was 221,376 ha from 2000 to 2016. Its impact to the hydrological system in the area was a key driver for the Philippine government to enact nationwide actions to address deforestation and degradation. The methodology of this thesis is capable of creating deforestation, reforestation, and climate change scenarios to assess the impacts of past government project i.e. NGP and support further decisions by the government.

Recommendations

Classifying subwatersheds as regulated or unregulated and including a separate grouping-learning method for this could deal with the mixed information given to the RF model. Otherwise, covariates indicating the presence of water regulating structures and dates of operation could be included.

Including sub-daily weather inputs together with lag time as one covariate could allow the RF model to learn the flood regulation service of forests. This "upgrade" could entail a trade-off between scale (decrease) and spatial-temporal resolution (increase). So, it could be tested in one typhoon-prone subwatershed.

The 8 weather inputs were the only non-spatial covariates for this study. A variogram for all weather stations in the Philippines could be created to spatially interpolate weather inputs using block kriging for two reasons: (1) come up with predicted spatial streamflow output; and (2) imputation of days with missing weather information. A high resolution spatial output of water services even at monthly temporal scale is advantageous in ecosystem accounting.

Yearly forest conversion pixels could be integrated to the RF model to better assess land conversion effect to seasonal water flows. Adding the land management practice given its data availability could make the conversion more realistic. This could also strengthen the forecasting method attempt in this thesis. Time series algorithms like BFAST could be used in the future. Sentinel 2 satellite images could be fused with other optical and even radar images for higher spatial-temporal resolution capturing forest degradation more. In line with the said possibilities, it would be better to test it in smaller scales to have more accurate land conversion classification in lesser processing time. Given a better "land conversion learning" by the RF model, ecosystem services could be depicted better by creating land cover scenarios. Additional associated covariates could be added relative to land conversion like evapotranspiration and base flow ratio. Lastly, a forest gain yearly pixel could be integrated especially now that the impact of the Philippine NGP needs to be assessed.

The results of this thesis could be supplemented to local land use planning. In the Philippines, the municipalities are mandated to enact their Forest Land Use Plans (FLUP) and Comprehensive Land Use Plans (CLUP). An effort to integrate ecosystem services accounting to those local plans was piloted in 3 sites by the DENR in 2016. The methodology and results of this study could help upscaling that effort.

This thesis' methodology could integrate with other modeling techniques for specific purposes. It could be used for flood modeling using process-based techniques to quantify the flood regulation service of forests. Moreover, it could be used for non-aggregated modeling techniques like Cellular Automate and Agent-based Models to include the social component (ecosystem stakeholders) in the modeling picture.

A more data driven approach to include every available subwatershed information, given that the methodology is almost fully automated, could be made. It could be more efficient if the weather and streamflow data can accessed by public online. A script that automatically integrates new subwatershed information could be made. This way, the model could learn other subwatersheds from areas outside the 6 river basins of this study. Otherwise, validation of the existing RF model outside Luzon could attest whether the current learned information by the RF model is sufficient for nationwide upscaling.

Lastly, a graphical user interface (GUI) using Shinny package in R could be made to make the methodology of this thesis more reproducible and user-friendly. The technical personnel of government offices like Climate Change Commission, DA, DENR, DOST, and DPWH could learn it and use this thesis methodology for their respective functions relative to ecosystem accounting and other similar programs, projects, activities (PPA).

References

Abon, C. C., David, C. P. C., & Pellejera, N. E. B. (2011). Reconstructing the tropical storm Ketsana flood event in Marikina River, Philippines. Hydrology and Earth System Sciences, 15(4), 1283.

Akhtar, M. K., Corzo, G. A., Van Andel, S. J., & Jonoski, A. (2009). River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges river basin. Hydrology and Earth System Sciences, 13(9), 1607.

Ali, S.A. (2015). Which DEM data is better for elevation mapping, ASTER or SRTM satellite data? accessed from https://researchgate.net.

Alibuyog, N. R., Ella, V. B., Reyes, M. R., Srinivasan, R., Heatwole, C., & Dillaha, T. (2009). Predicting the effects of land use change on runoff and sediment yield in Manupali River subwatersheds using the SWAT model. International Agricultural Engineering Journal, 18(1), 15.

Aquino, D., Paningbatan, E., & Mahar Francisco Lagmay, A. (2014, May). Predicting the Discharge Rate Contribution of the Binuwang Watershed to the Agos River, Philippines. In EGU General Assembly Conference Abstracts (Vol. 16).

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., & Kannan, N. (2012). SWAT: Model use, calibration, and validation. Transactions of the ASABE, 55(4), 1491-1508.

Ataroff, M., & Rada, F. (2000). Deforestation impact on water dynamics in a Venezuelan Andean cloud forest. AMBIO: A Journal of the Human Environment, 29(7), 440-444.

Baguinon, N. T., Lasco, R. D., Macandog, D. M., Pasicolan, P. N., & Villancio, V. T. (2007). Agroforestry and land use in the Philippines. World Agroforestry, Philippines, 108-127.

Bateman, I. J., Harwood, A. R., Mace, G. M., Watson, R. T., Abson, D. J., Andrews, B., ... & Fezzi, C. (2013). Bringing ecosystem services into economic decision-making: land use in the United Kingdom. science, 341(6141), 45-50.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Broich, M., Hansen, M. C., Potapov, P., Adusei, B., Lindquist, E., & Stehman, S. V. (2011). Time-series analysis of multi-resolution optical imagery for quantifying forest cover loss in Sumatra and Kalimantan, Indonesia. International Journal of Applied Earth Observation and Geoinformation, 13(2), 277-291.

Bureau of Soils and Water Management (undated). Irrigation of small-scale farm lands in the Philippines. http://www.bswm.da.gov.ph/.

Calder, I. R., & Aylward, B. (2006). Forest and floods: Moving to an evidence-based approach to watershed and integrated flood management. Water International, 31(1), 87-99.

Calder, I., Hofer, T., Vermont, S., & Warren, P. (2008). Towards a new understanding of forests and water. UNASYLVA-FAO-, 229, 3.

Carlisle, D. M., Wolock, D. M., & Meador, M. R. (2011). Alteration of streamflow magnitudes and potential ecological consequences: a multiregional assessment. Frontiers in Ecology and the Environment, 9(5), 264-270.

Christiansen, D. E. (2012). Simulation of daily streamflows at gaged and ungaged locations within the Cedar River Basin, Iowa, using a Precipitation-Runoff Modeling System model (No. 2012-5213). US Geological Survey.

Cruz, R. V. O., AliÃśo, P. M., Cabrera O. C., David, C. P. C., David, L. T., Lansigan, F. P., Lasco, R. D., Licuanan, W. R. Y., Lorenzo, F. M., Mamauag, S. S., PeÃśaflor, E. L., Perez, R. T., Pulhin, J. M., Rollon, R. N., Samson, M. S., Siringan, F. P., Tibig, L. V., Uy, N. M., Villanoy, C. L. (2017). 2017 Philippine Climate Change Assessment: Impacts, Vulnerabilities and Adaptation. The Oscar M. Lopez Center for Climate Change Adaptation and Disaster Risk Management Foundation, Inc. and Climate Change Commission.

de Araujo Barbosa, C. C., Atkinson, P. M., & Dearing, J. A. (2015). Remote sensing of ecosystem services: a systematic review. Ecological Indicators, 52, 430-443.

Department of Budget and Management (undated). 2016 Budget Allocation to Infrastructures. https://www.dbm.gov.ph/.

Department of Environment and Natural Resources-Forest Management Bureau (undated). Philippine Forestry Statistics. http://forestry.denr.gov.ph/index.php/statistics/philippines-forestry-statistics.

Department of Public Works and Highways (undated). Flood control structures report. http://www.dpwh.gov.ph/dpwh/news/.

Engman, E. T. (1986). Roughness coefficients for routing surface runoff. Journal of Irrigation and Drainage Engineering, 112(1), 39-53.

Erwin, Martha L., and Pixie A. Hamilton. Monitoring Our Rivers and Streams. US Department of the Interior, US Geological Survey, (2002). APA .

Feng, X., Fu, B., Yang, X., & LÃij, Y. (2010). Remote sensing of ecosystem services: An opportunity for spatially explicit assessment. Chinese Geographical Science, 20(6), 522-535.

Galelli, S., & Castelletti, A. (2013). Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. Hydrology and Earth System Sciences, 17(7), 2669.

Goldstein, J. H., Caldarone, G., Duarte, T. K., Ennaanay, D., Hannahs, N., Mendoza, G., ...& Daily, G. C. (2012). Integrating ecosystem-service tradeoffs into land-use decisions. Proceedings of the National Academy of Sciences, 109(19), 7565-7570.

Granato, G. E. (2012). Estimating basin lagtime and hydrograph-timing indexes used to characterize stormflows for runoff-quality analysis. Scientific Investigations Report, 5110.

Grimaldi, S., Petroselli, A., & Romano, N. (2013). GreenâĂŘAmpt CurveâĂŘNumber mixed procedure as an empirical tool for rainfallâĂŞrunoff modelling in small and ungauged basins. Hydrological processes, 27(8), 1253-1264.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S., Tyukavina, A., & Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. science, 342(6160), 850-853.

Hurkmans, R. T. W. L., Terink, W., Uijlenhoet, R., Moors, E. J., Troch, P. A., & Verburg, P. H. (2009). Effects of land use changes on streamflow generation in the Rhine basin. Water resources research, 45(6).

Jones, J. A., & Grant, G. E. (1996). Peak flow responses to clearâĂŘcutting and roads in small and large basins, western Cascades, Oregon. Water Resources Research, 32(4), 959-974.

Jones, J. A., Achterman, G. L., Augustine, L. A., Creed, I. F., Ffolliott, P. F., MacDonald, L., & Wemple, B. C. (2009). Hydrologic effects of a changing forested landscapeâĂŤchallenges for the hydrological sciences. Hydrological Processes, 23(18), 2699-2704.

Klepeis, P., Scull, P., LaLonde, T., Svajlenka, N., & Gill, N. (2013). Changing forest recovery dynamics in the northeastern United States. Area, 45(2), 239-248.

Koch, F. J., Van Griensven, A., Uhlenbrook, S., Tekleab, S., & Teferi, E. (2012). The Effects of land use change on hydrological responses in the choke mountain range (Ethiopia)-a new approach addressing land use dynamics in the model SWAT.

Koschke, L., FÃijrst, C., Frank, S., & Makeschin, F. (2012). A multi-criteria approach for an integrated land-cover-based assessment of ecosystem services provision to support landscape planning. Ecological Indicators, 21, 54-66.

Kroll, C., Luz, J., Allen, B., & Vogel, R. M. (2004). Developing a watershed characteristics database to improve low streamflow prediction. Journal of Hydrologic Engineering, 9(2), 116-125.

Kummer, D. M. (1992). Deforestation in the postwar Philippines (No. 234). University of Chicago Press.

Levin, S., Xepapadeas, T., CrÃľpin, A. S., Norberg, J., De Zeeuw, A., Folke, C., ... & Ehrlich, P. (2013). Social-ecological systems as complex adaptive systems: modeling and policy implications. Environment and Development Economics, 18(2), 111-132.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Maes, J., Teller, A., Erhard, M., Liquete, C., Braat, L., Berry, P., & Paracchini, M. L. (2013). Mapping and assessment of ecosystems and their services-An analytical framework for ecosystem assessments under action 5 of the EU biodiversity strategy to 2020.

Manarang, M. N. S. J., Ho H.T. & Anicete, A. B. (2000). Increased efficiency in vetiver propagation with the use of growth promoters. Proc. ICV-2, ORDPB.

Milly, Paul CD, Kathryn A. Dunne, and Aldo V. Vecchia. "Global pattern of trends in streamflow and water availability in a changing climate." Nature 438.7066 (2005): 347-350.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3), 885-900.

Mustard, John F., and Thomas R. Fisher. "Land use and hydrology." Land Change Science. Springer Netherlands, 2012. 257-276.

Mustard, J. F., & Fisher, T. R. (2012). Land use and hydrology. In Land Change Science (pp. 257-276). Springer, Dordrecht.

National Irrigation Authority (2011). "Small Reservoir Irrigation Project Impact's Farmers Lives". http://www.nia.gov.ph/?q=content/small-reservoir-irrigation-project-impacts-farmers-lives.

National Disaster Risk Reduction and Management Office (undated). DRRM Knowledge Center. http://drrmkc.teamyap.com/.

National Mapping and Resource Information Authority (undated). Philippine Land Cover Classification Report. http://www.namria.gov.ph/products.aspx.

Nelson, E., Mendoza, G., Regetz, J., Polasky, S., Tallis, H., Cameron, D., ... & Lonsdorf, E. (2009). Modeling multiple ecosystem services, biodiversity conservation, commodity production, and tradeoffs at landscape scales. Frontiers in Ecology and the Environment, 7(1), 4-11.

Ogden, F. L., Crouch, T. D., Stallard, R. F., & Hall, J. S. (2013). Effect of land cover and use on dry season river runoff, runoff efficiency, and peak storm runoff in the seasonal tropics of Central Panama. Water Resources Research, 49(12), 8443-8462.

Pasquarella, V. J., Holden, C. E., Kaufman, L., & Woodcock, C. E. (2016). From imagery to ecology: Leveraging time series of all available Landsat observations to map and monitor ecosystem state and dynamics. Remote Sensing in Ecology and Conservation, 2(3), 152-170.

Philippine Atmospheric Geophysical and Astronomical Service Administration (2016)."Climate change scenarios using PRECIS". https://www1.pagasa.dost.gov.ph/index.php/93-cad1/470-climate-change-scenarios.

Philippine Statistics Authority (undated). "Selected statistics on Agriculture". accessed from https://psa.gov.
selected-statistics-agriculture.

R Documentation (undated). Package set.seed description. https://www.rdocumentation.org/packages/simEd/versions/1.0.3/topics/set.seed.

Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. Journal of Hydrology, 414, 284-293.

Rawlins M., Aggabao LF., Araza A., Calderon M., Elomina J., Ignacio.,. & Soyosa E. (2017) "Understanding the Role of Forests in Supporting Livelihoods and Climate Resilience: Case Studies in the Philippines". World Bank: Manila, Philippines.

River Basin Coordinating Council (undated). "River Basin Profiles". accessed from https://rbco.denr.gov.ph.

Sahin, V., & Hall, M. J. (1996). The effects of afforestation and deforestation on water yields. Journal of hydrology, 178(1-4), 293-309.

Sando, R., & Chase, K. J. (2017). Estimating current and future streamflow characteristics at ungaged sites, central and eastern Montana, with application to evaluating effects of climate change on fish populations (No. 2017-5002). US Geological Survey.

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. Hydrology and Earth System Sciences, 20(7), 2611.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. BMC bioinformatics, 9(1), 307.

Tong, H. (2012). Threshold models in non-linear time series analysis (Vol. 21). Springer Science & Business Media.

Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. Nature, 540(7633), 418-422.

Pilot Ecosystem Account for Southern Palawan (2017). accessed from https://wavespartnership.org. Accessed 23 Oct.

Principe, J. A. (2012). Exploring climate change effects on watershed sediment yield and land cover-based mitigation measures using swat model, RS and GIS: case of Cagayan River Basin, Philippines. Int. Arch. Photogram. Rem. Sens. Spatial Inform. Sci, 39, 193-198.

Remme, R. P., SchrÃűter, M., & Hein, L. (2014). Developing spatial biophysical accounting for multiple ecosystem services. Ecosystem Services, 10, 6-18.

Reynolds, L. V., & Shafroth, P. B. (2016). Modeled streamflow metrics on small, ungaged stream reaches in the Upper Colorado River Basin (No. 974). US Geological Survey.

Rodriguez-MartÃŋnez, J., & Santiago, M. (2017). The effects of forest cover on base flow of streams in the mountainous interior of Puerto Rico, 2010 (No. 2016-5142). US Geological Survey.

Silent Garden (undated). Typhoons in the Philippines. acquired from: https://www.silent-gar

dens.com/climate.php.

Sukhdev, P. (2008). The economics of ecosystems and biodiversity. European Communities, 3.

Sumatiphala, S. (2014). How to determine the number of trees to be generated in Random Forest algorithm? accessed from https://researchgate.net.

Swetnam, R. D., Fisher, B., Mbilinyi, B. P., Munishi, P. K., Willcock, S., Ricketts, T., ... & Lewis, S. L. (2011). Mapping socio-economic scenarios of land cover change: A GIS method to enable ecosystem service modelling. Journal of environmental management, 92(3), 563-574.

Van Remortel, R. D., Hamilton, M. E., & Hickey, R. J. (2001). Estimating the LS factor for RUSLE through iterative slope length processing of digital elevation data within Arclnfo grid. Cartography, 30(1), 27-35.

Vargas, L., Hein, L., & Remme, R. P. (2017). Accounting for ecosystem assets using remote sensing in the Colombian Orinoco River Basin lowlands. Journal of Applied Remote Sensing, 11(2), 026008-026008.

Verbesselt, J., Zeileis, A., & Herold, M. (2012). Near real-time disturbance detection using satellite image time series. Remote Sensing of Environment, 123, 98-108.

Volante, J. N., Alcaraz-Segura, D., Mosciaro, M. J., Viglizzo, E. F., & Paruelo, J. M. (2012). Ecosystem functional changes associated with land clearing in NW Argentina. Agriculture, Ecosystems & Environment, 154, 12-22.

Wang, W. (2006). Stochasticity, nonlinearity and forecasting of streamflow processes. Ios Press.

Wang, L., & Ren, Z. (2008). Spatial-temporal differences in in-stream flow requirement based on GIS: A case study of YanâĂŹan region, northern Shaanxi. Journal of Geographical Sciences, 18(1), 107-114.

Winiger, M. G. H. Y., Gumpert, M., & Yamout, H. (2005). KarakorumâĂŞHindukushâĂŞwestern Himalaya: assessing highâĂŘaltitude water resources. Hydrological Processes, 19(12), 2329-2338.

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

Xu, L., Saatchi, S. S., Yang, Y., Yu, Y., & White, L. (2016). Performance of non-parametric algorithms for spatial mapping of tropical forest structure. Carbon balance and management, 11(1), 18.

Zhao, T., Yang, D., Cai, X., & Cao, Y. (2012). Predict seasonal low flows in the upper Yangtze River using random forests model. Journal of Hydroelectric Engineering, 3(005).

Zhao, L., Duan, Q., Schaake, J., Ye, A., & Xia, J. (2011). A hydrologic post-processor for ensemble streamflow predictions. Advances in Geosciences, 29, 51-59.

Zhou, G. Y., Morris, J. D., Yan, J. H., Yu, Z. Y., & Peng, S. L. (2002). Hydrological impacts of

reafforestation with eucalypts and indigenous species: a case study in southern China. Forest Ecology and Management, 167(1-3), 209-222.

A Acronym and actual name of subwatersheds.

Subwatershed code	Basin	Subwatershed name	Area (ha)
aarb_a	Apayao-Abulug River Basin	Abulug	236,119
aarb_n	Apayao-Abulug River Basin	Nagan	19,277
abrb_s	Abra River Basin	Sinalang	1,461
arb_b	Agno River Basin	Bayaoas	31,609
arb_c	Agno River Basin	Camiling	36,507
crb_a	Cagayan River Basin	Aligapay	104,525
crb_be	Cagayan River Basin	Benay	27,211
crb_bu	Cagayan River Basin	Buenavista	1,851,853
crb_d	Cagayan River Basin	Dipantan	169,590
crb_j	Cagayan River Basin	Jones	293,539
crb_m	Cagayan River Basin	Magat	215,601
crb_p	Cagayan River Basin	Pinacuanan	1,187,380
crb_s	Cagayan River Basin	Saltan	85,267
crb_t	Cagayan River Basin	Pinacuanan Tugue	66,994
crb_u	Cagayan River Basin	Upi	118,738
mrb_s	Marikina River Basin	San Jose	63,445
prb_a	Pampanga River Basin	Arayat	622,280
prb_b	Pampanga River Basin	Baliwag	37,406
prb_c	Pampanga River Basin	Coronel	84,287
prb_p	Pampanga River Basin	Poblacion	85,739
prb_r	Pampanga River Basin	Rio Chico	147,804

B Scripts used to pre-process covariates.

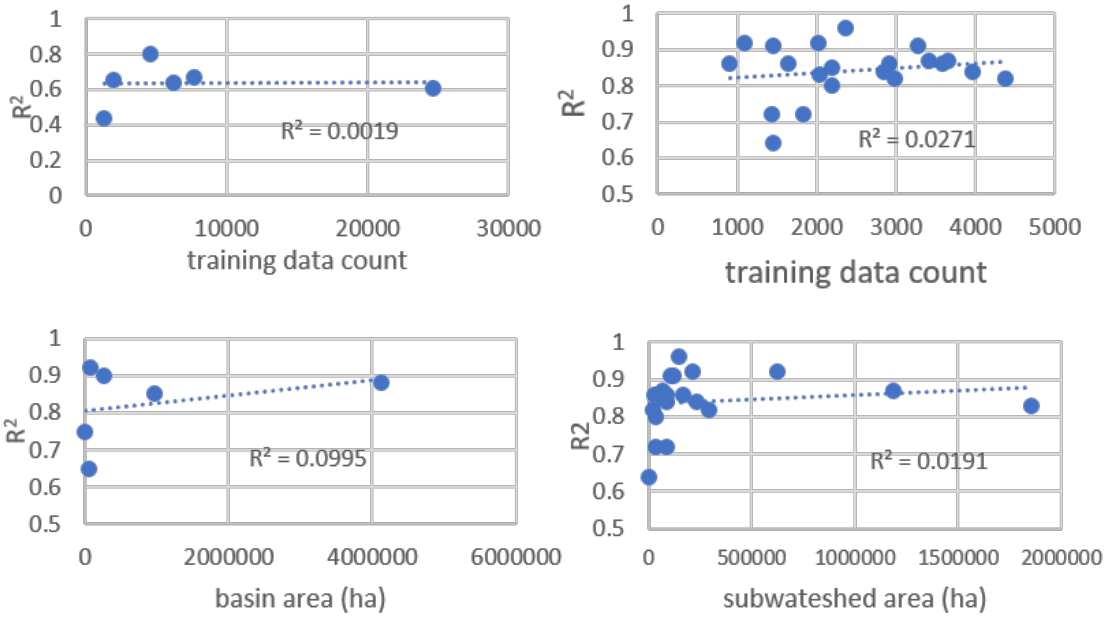
```
## SCRIPT TO PRE-PROCESS DEM AND SLOPE INPUTS

# Preliminaries
pacman::p_load (raster, rgdal)
mydir <- setwd('M:/THESIS_PP')

# Function to open files from subfolders: general, weather, soil, lcover, dem
FolderFiles <- function(folder, ext) {
  allfiles <- list.files(paste0(mydir,'/data/', folder),  pattern = paste0('*.', ext, '$'))
  print (paste('loaded:', allfiles))
  return (allfiles)
}

# Read loss-gain pixels, mask, and write
DEMfiles <- FolderFiles ('dem', 'tif')

DEM <- raster(paste0(mydir,'/data/dem/', DEMfiles[2]))
slope <- raster(paste0(mydir,'/data/dem/', DEMfiles[3]))
DEM.slope <- stack(DEM, slope)

dshape <- readOGR(dsn = 'M:/THESIS_PP/data/general', layer = "subs1")
dcrop <- crop(DEM.slope, extent(shape))
dmask <- mask(dcrop, shape)

setwd('M:/THESIS_PP/mid-results')
writeRaster(dmask, filename="DEM.sl.tif", overwrite=TRUE)

```

```
## SCRIPT TO PRE-PROCESS FOREST LOSS AND LAND COVER INPUTS

# Preliminaries
pacman::p_load (raster, rgdal)
mydir <- setwd('D:/THESIS_PP')

# Function to open files from subfolders: general, weather, soil, lcover, dem
FolderFiles <- function(folder, ext) {
  allfiles <- list.files(paste0(mydir,'/data/', folder),  pattern = paste0('*.', ext, '$'))
  print (allfiles)
}

# Read land cover and loss-gain pixels then set extent, mask and write
lcovtifs <- FolderFiles ('lcover', 'tif')
floss <- raster(paste0(mydir,'/data/lcover/', as.character(lcovtifs)[2]))
fgain <- raster(paste0(mydir,'/data/lcover/', as.character(lcovtifs)[1]))
lcov1 <- raster(paste0(mydir,'/data/lcover/', as.character(lcovtifs)[5]))
lcov2 <- raster(paste0(mydir,'/data/lcover/', as.character(lcovtifs)[6]))
lcov1 <- setExtent(lcov1, lcov2)
lcov1 [lcov1==0] <- NA
lcov2 [lcov2==0] <- NA
lcov1 [lcov1==255] <- NA
lcov2 [lcov2==255] <- NA
NetLoss <- mask(floss, fgain)

lshape <- readOGR(dsn = 'D:/THESIS_PP/data/general', layer = "subs1")
lcrop1 <- crop(lcov1, extent(lshape))
lcrop2 <- crop(lcov2, extent(lshape))
lmask1 <- mask(lcrop1, lshape)
lmask2 <- mask(lcrop2, lshape)
lmask1 <- setExtent(lmask1, lmask2)
lmask1 [lmask1 == 0] <- NA
lmask2 [lmask2 == 0] <- NA
lmask2 [lmask2 == 255] <- NA
lcovs <- list(lcov1, lcov2)
lcovs$fun <- mean
lcov2000 <- do.call(mosaic, lcovs)

setwd('D:/THESIS_PP/mid-results')
writeRaster(lmask, filename="Lcov00.tif", overwrite=TRUE)
hist(lmask)

```

```
### Script to impute missing weather data ###

# Preliminaries - packages, directory
pacman::p_load(imputeTS, Hmisc, rio, plyr, dplyr, xlsx)
mydir <- setwd ('D:/THESIS_PP')

# Load weather data in xls and converts to csv
loadcsv <- function(datadir){
  setwd (datadir)
  files.to.read = list.files(datadir, pattern="xls")
  lapply(files.to.read, function(f) {
    df = read.xlsx(f, sheetIndex=1)
    write.csv(df, gsub("xls", "csv", f), row.names=FALSE)
    setwd(mydir)
  })}

allcsv <- loadcsv('H:/THESIS_PP/data/weather/updated/')

# Function to fill small gaps via interpolation; if NAs are smaller than given 'gap', the funciton will interpolat
FillSmall <- function (datadir, gap) {
  setwd (datadir)
  alltables <- list.files (datadir, pattern='*.csv')
  alltables0 <- lapply(alltables, read.csv)
  alltables1 <- lapply(alltables0, "[", TRUE, -1) # delete first column, not needed for regression
  alltables2 <- lapply(alltables1, function(x) x[-c(6941:10000),])# delete excess rows
  alltables3 <- bind_cols(alltables2) # won't bind if different rowsums
  alltables3 <- alltables3[, colSums(is.na(alltables3)) != nrow(alltables3)]


  for (col in names(alltables3)) {
    missing <- sum(is.na(alltables3[,col]))
    if (missing <= gap & missing != 0) {
      fill <- na.interpolation(alltables3[,col], option='spline')
      alltables3[,col] <- fill
      setwd('D:/THESIS_PP/mid-results/')
      write.csv (alltables3, file='int_weather1.csv')
      }
  }
    if (missing >= length(col)){
      col_pos <- which(colnames(alltables3)==col)
      alltables3 <- alltables3[,-col_pos]
  }
  setwd('D:/THESIS_PP/mid-results/')
 # alltables3 <-  alltables3[, -which(colMeans(is.na(alltables3)) > 1o)]
  write.csv (alltables3, file='int_weather1.csv')
  return(alltables3)
}
df <- FillSmall('D:/THESIS_PP/data/weather/updated/',50)
df.x <- df[,colSums(is.na(df))<nrow(df)]

# Function to fill large gaps
NoMissing <- function (gap){
  w.table <- read.csv('D:/THESIS_PP/mid-results/int_weather1.csv')
  w.table <- w.table[,-1]
  w.table <- w.table[,colSums(is.na(w.table))<nrow(w.table)]
missing <- sum(is.na(w.table[,col]))
#if (sum(!is.na(col)) < 2 ) {return(NA)}
if (missing <= gap) {
  col_pos <- which(colnames(w.table)==col)
  ind <- w.table[,-(as.numeric(col_pos))]
  dep <- w.table[,col]
  coi <- w.table[[col_pos]]

  # first batch regression removing column of interest
  reg <- lapply(ind, function(x) (summary(lm(dep ~ x))))

  # get highest R2
  r2 <- ldply(reg, function(x) {r.sq <- x[1:length(x)]$r.squared
  data.frame(r.sq)})
  r2high <- order(-r2$r.sq)
  r2high <- r2high[1]
  predname <- noquote(r2$.id[r2high])
  pred <- which (colnames(w.table)==predname)

  # regress again and predict using highest R2
  result <- lm((dep) ~ (w.table[[pred]]))
  newpred <- predict(result, newdata= data.frame(dep))

  # retain original values, replace with predicted column, and write new table
  is.na(w.table[[col_pos]][]) <- setNames(w.table[[col_pos]], newpred)[as.character(unlist(is.na(w.table[[col_pos]])))]
  w.table[[col_pos]] <- newpred
  indx <- !is.na(coi)
  w.table[[col_pos]][indx] <- coi[indx]
  setwd('D:/THESIS_PP/mid-results/')
  write.csv(w.table, 'reg_weather1.csv')
  setwd(mydir)
 }
}
for (i in names (w.table)) {print(sum(is.na((w.table[,i]))))} #checker
return(w.table)
```

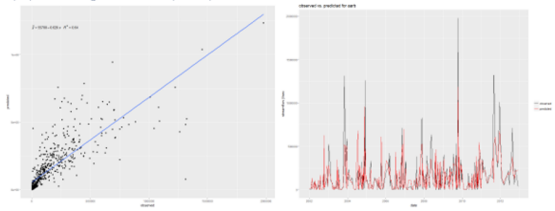
```
## SCRIPT TO PRE-PROCESS SOIL INPUTS

# Preliminaries
pacman::p_load (raster, rgdal)
mydir <- setwd('D:/THESIS_PP')

# Function to open files from subfolders: general, weather, soil, lcover
FolderFiles <- function(folder, ext) {
  allfiles <- list.files(paste0(mydir,'/data/', folder),  pattern = paste0('*.', ext, '$'))
  print (allfiles)
}

soiltifs <- FolderFiles ('soil', 'tif')
print (soiltifs)

# Function to read tifs, mosaic and mask soil inputs
ReadSoil <- function (soilcode){
  index <- grep(soiltifs,  pattern = as.character(paste0('^', soilcode)))
  input <- lapply(index, function(x) raster(paste0('M:/THESIS_PP/data/soil/', as.character(soiltifs[x]))))
  input$fun <- mean
  smosaic <- do.call(mosaic, input)
  shape <- readOGR(dsn = 'M:/THESIS_PP/data/general', layer = "subs1")
  sproj <- projectRaster(smosaic, crs = shape@proj4string)
  scrop <- crop(sproj, extent(shape))
  smask <- mask(scrop, shape)
  return (smask)
}

SoilClass <- ReadSoil ('TAXNWRB')
Bedrock <- ReadSoil ('BDTICM')
BulkDensity <- ReadSoil ('BLDFIE')
Sand <- ReadSoil ('SNDPPT')
Silt <- ReadSoil ('SLTPPT')
Clay <- ReadSoil ('CLYPPT')


setwd('M:/THESIS_PP/mid-results')
writeRaster(SoilClass, filename="SoilClass.tif", overwrite=TRUE)
writeRaster(Bedrock, filename="Bedrock.tif", overwrite=TRUE)
writeRaster(BulkDensity, filename="BulkDensity.tif", overwrite=TRUE)
writeRaster(Sand, filename="Sand.tif", overwrite=TRUE)
writeRaster(Clay, filename="Clay.tif", overwrite=TRUE)
writeRaster(Silt, filename="Silt.tif", overwrite=TRUE)
setwd('M:/THESIS_PP')
```

C Regression results of accuracy measure to area and number of observations.

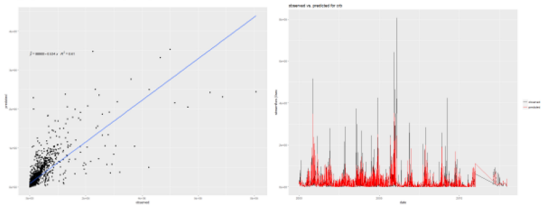


D Basin and subwatershed models validation results.

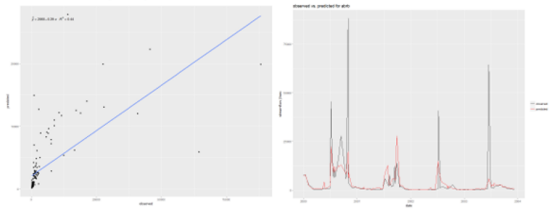
Apayao-Abulug River Basin (AARB)



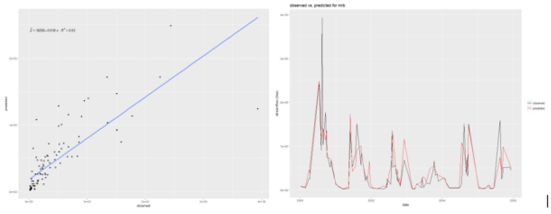
Cagayan River Basin (CRB)



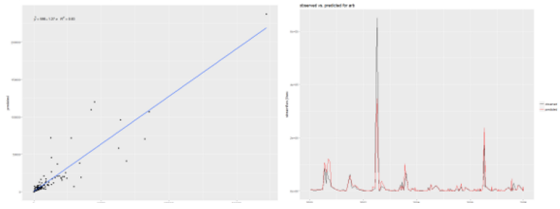
Abra River Basin (ABRB)



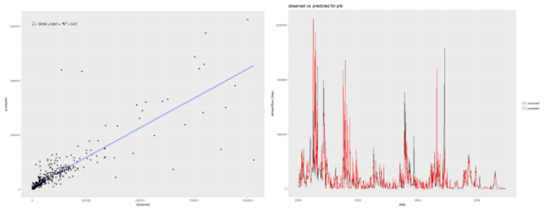
Marikina River Basin (MRB)



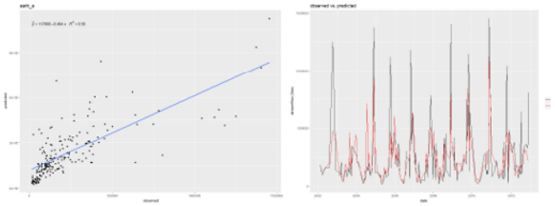
Agno River Basin (ARB)



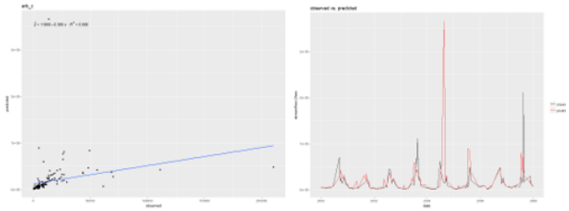
Pampanga River Basin (PRB)



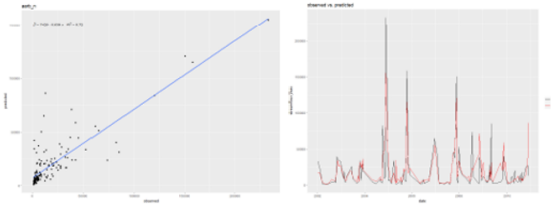
aarb_a



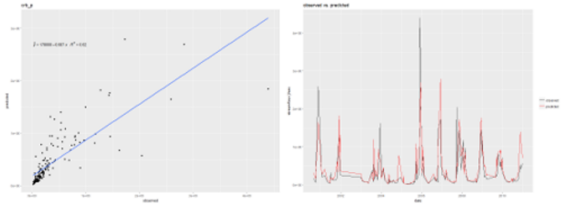
arb_c



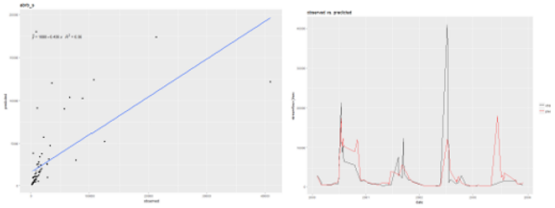
abrb_n



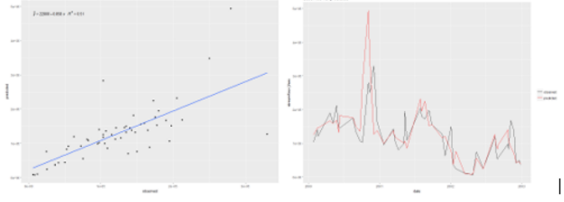
crb_p



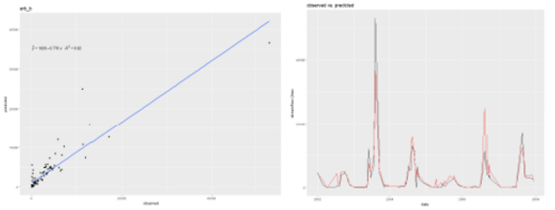
abrb_s



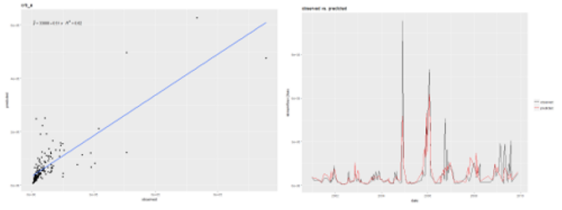
crb_m



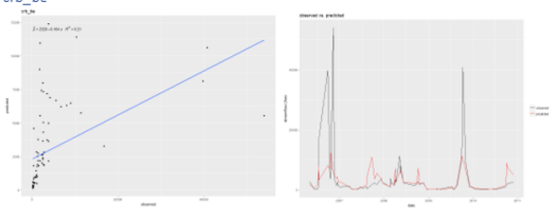
arb_b



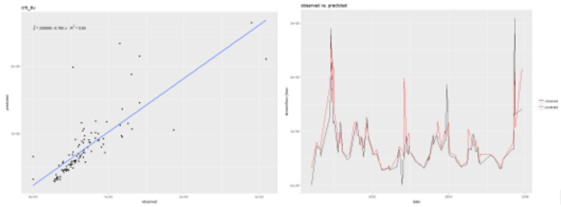
crb_a



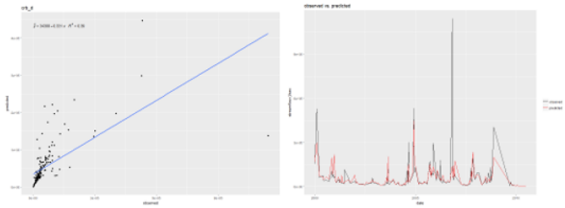
crb_be



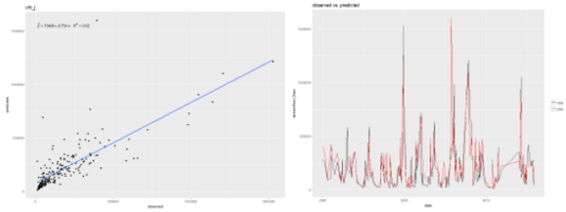
crb_bu



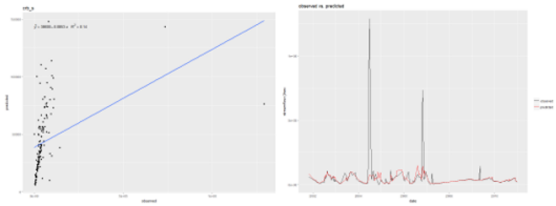
crb_d



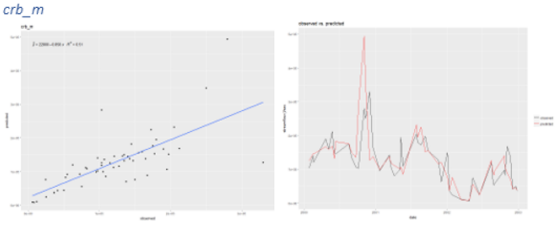
crb_j



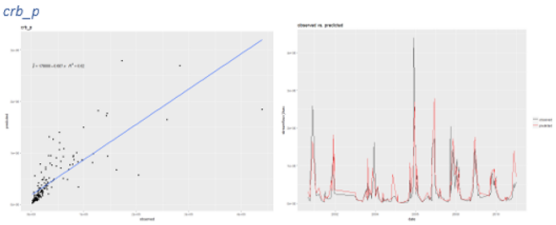
crb_s



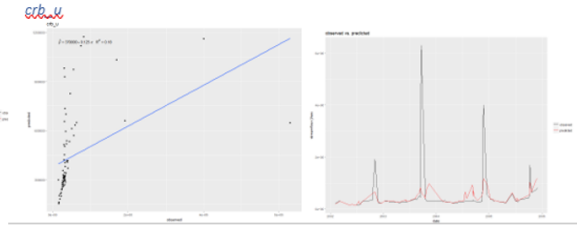
crb_m



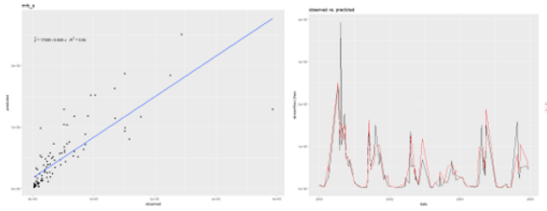
crb_p



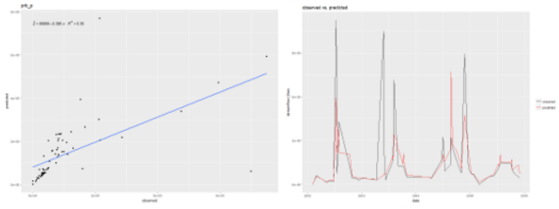
crb_w



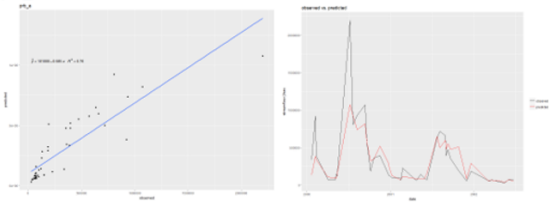
mrb_s



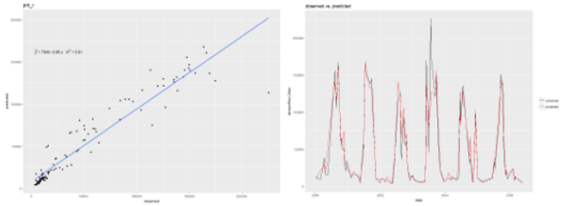
prb_p



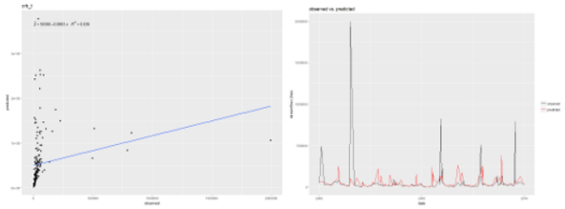
prb_a



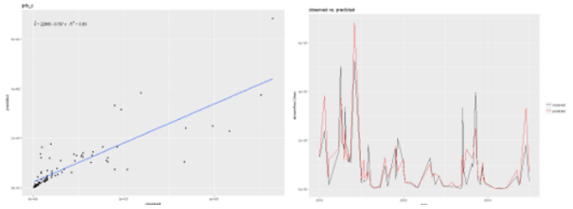
prb_r



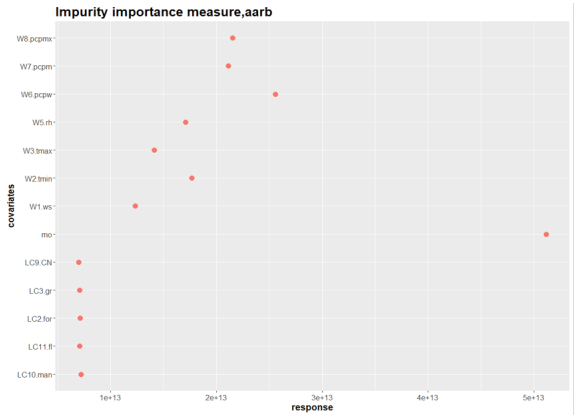
crb_t

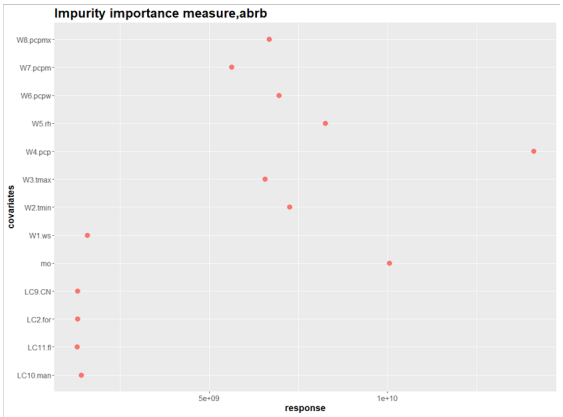


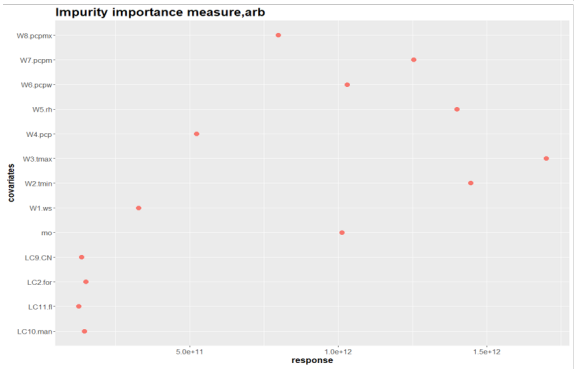
prb_c

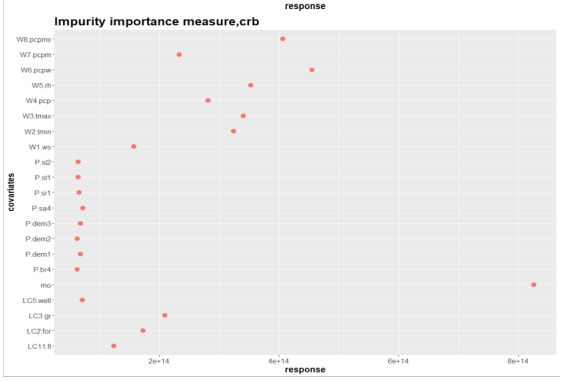


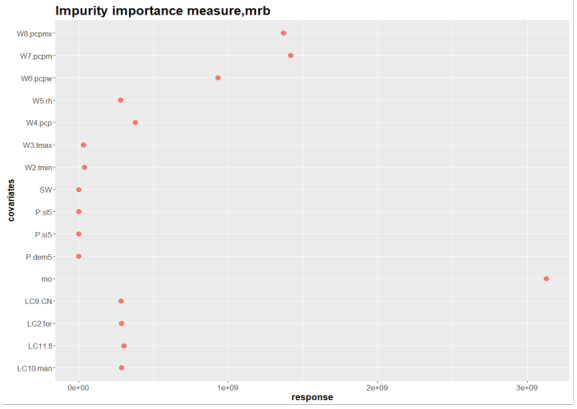
E Variable importance measure graphs for basins and sub-watersheds.

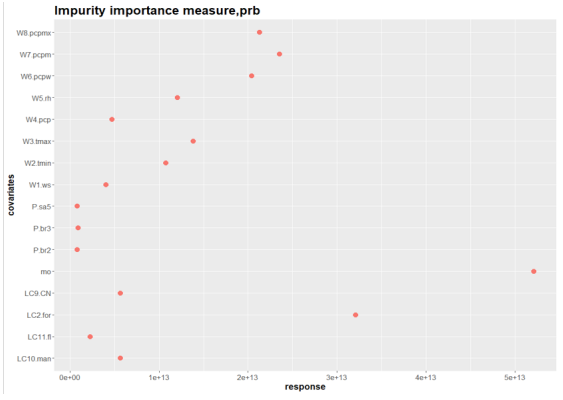


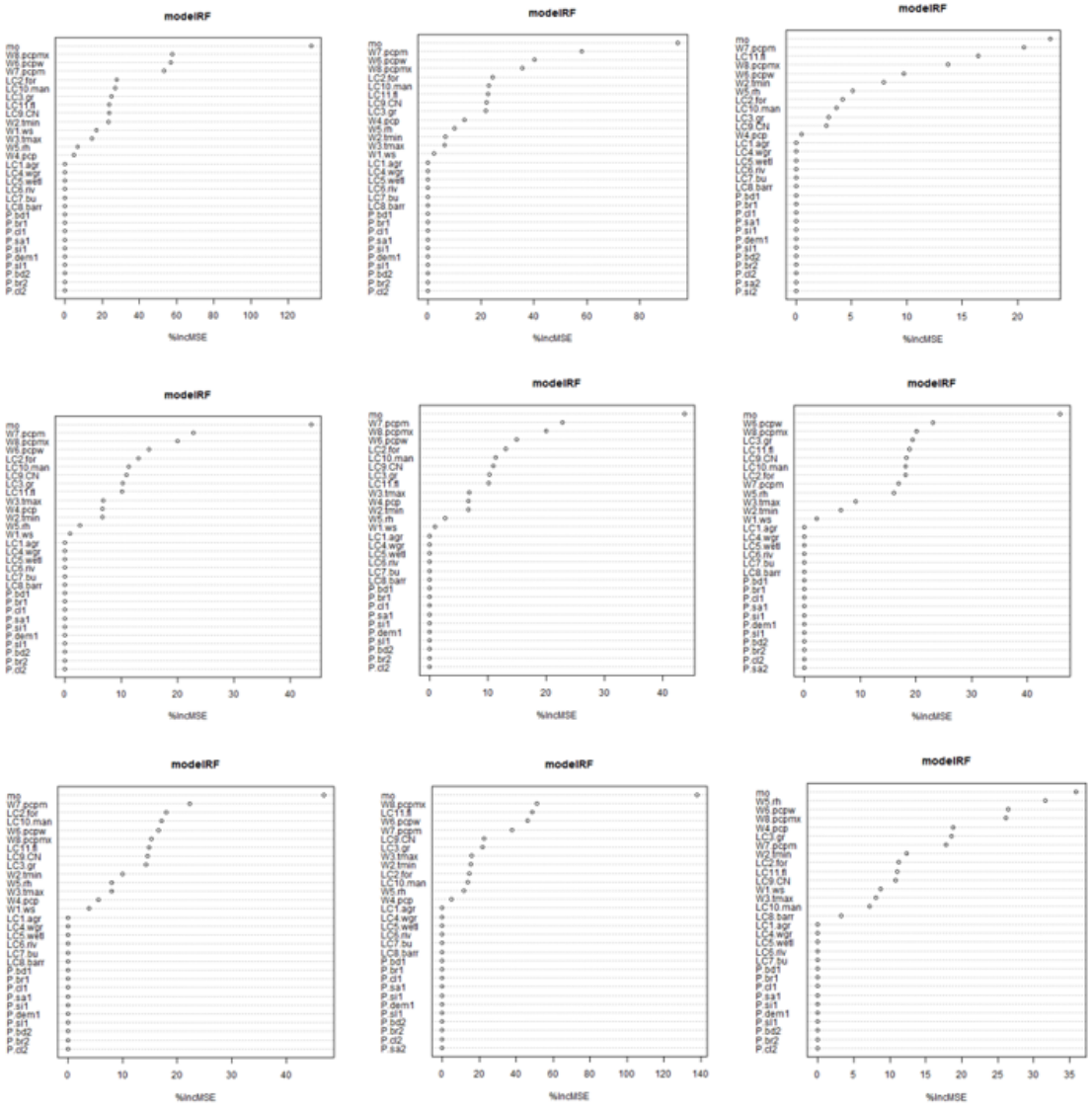


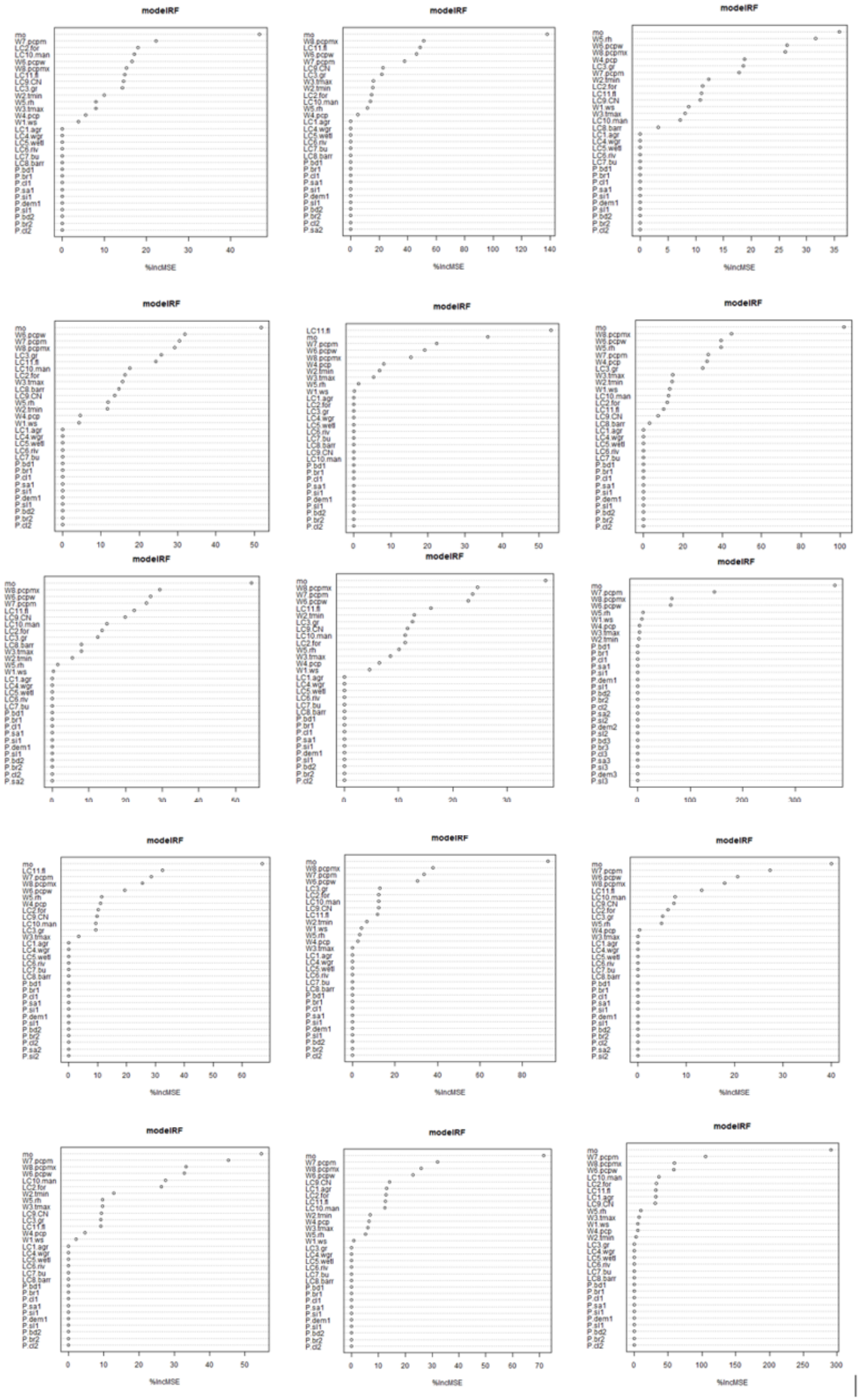




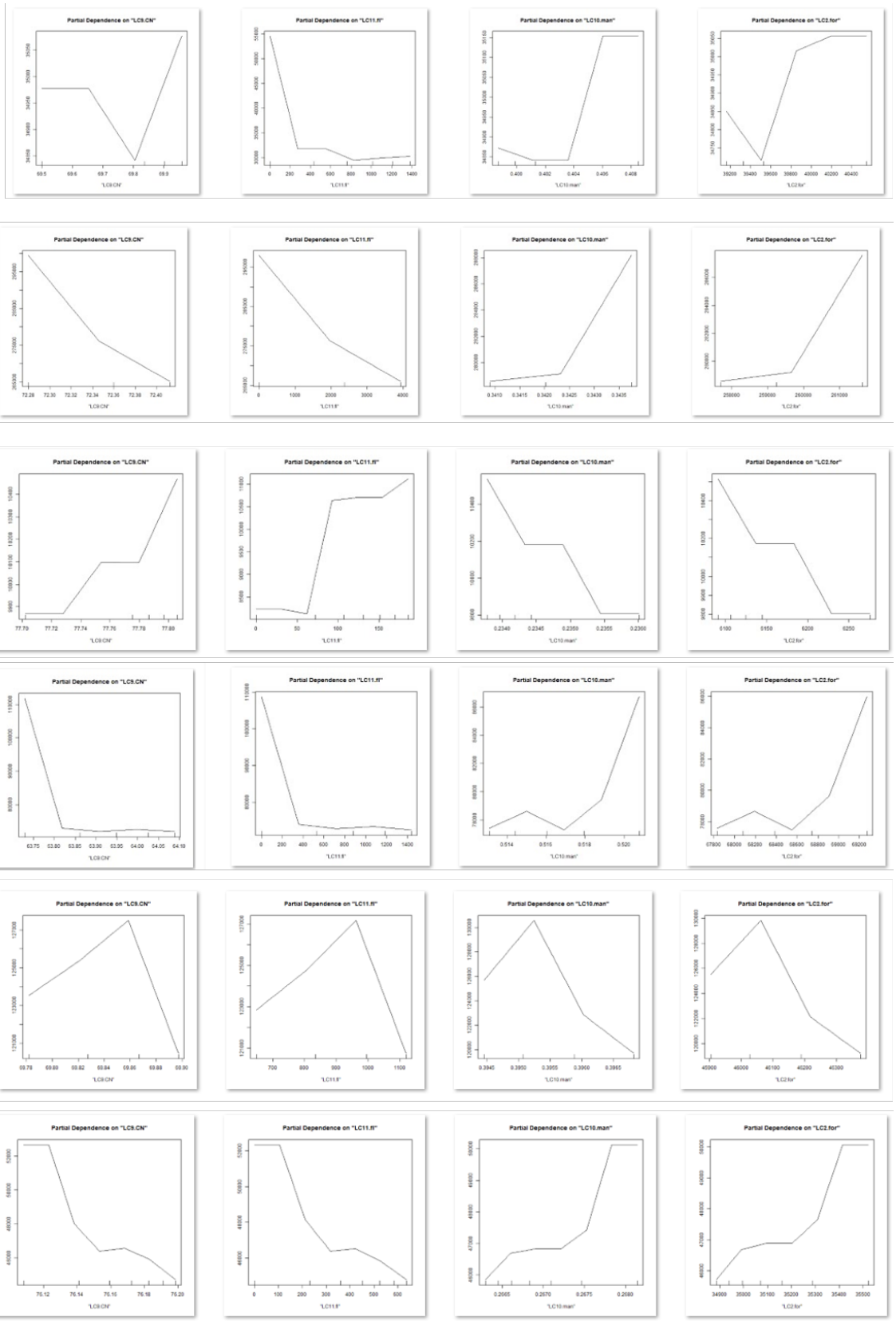


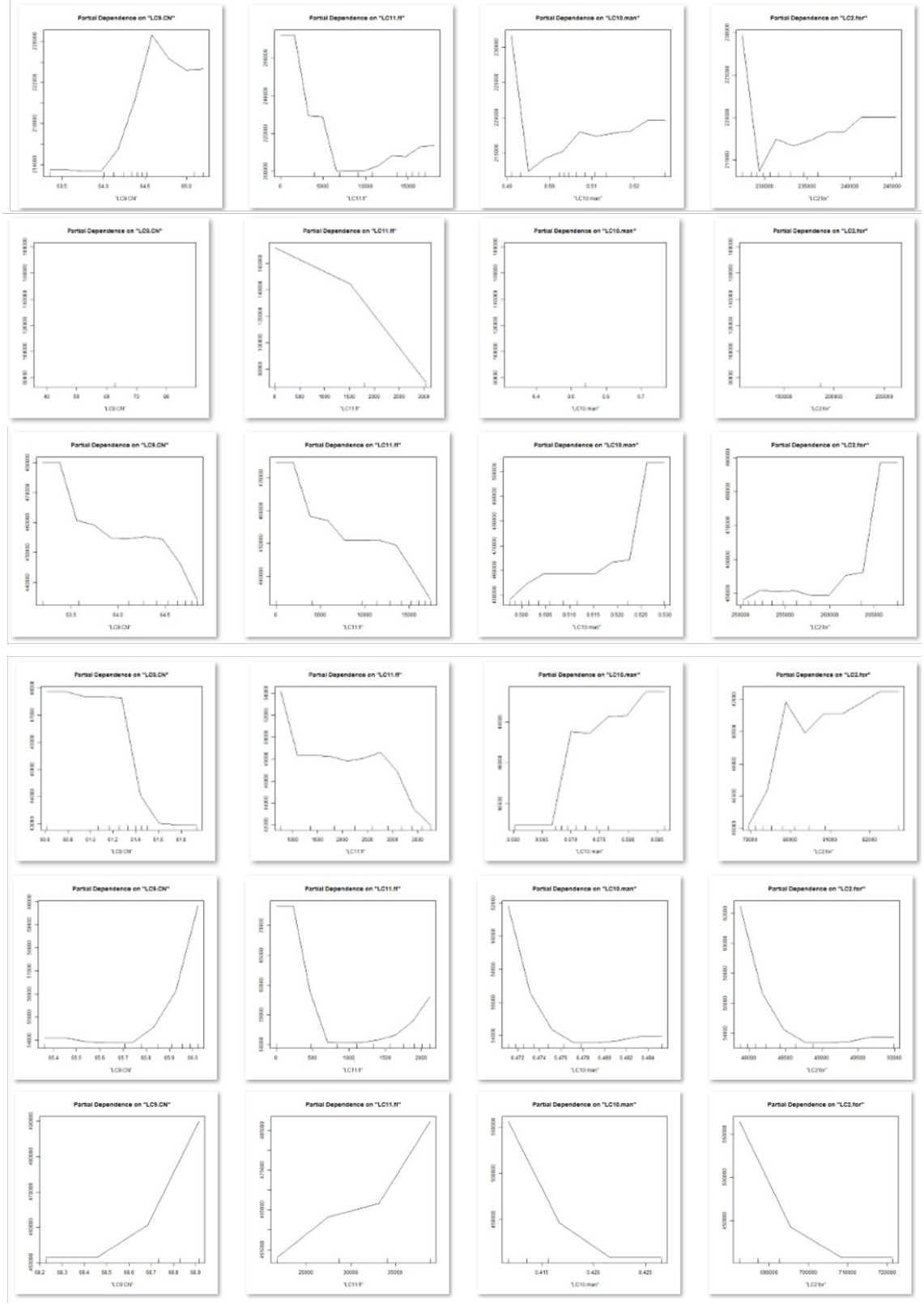



modelRF
%IncMSE
modelRF
%IncMSE
modelRF
%IncMSE
modelRF
%IncMSE
modelRF
%IncMSE
modelRF
%IncMSE
modelRF
%IncMSE
modelRF
%IncMSE
modelRF
%IncMSE


modelRF
%IncMSE

F PDP per subwatershed.

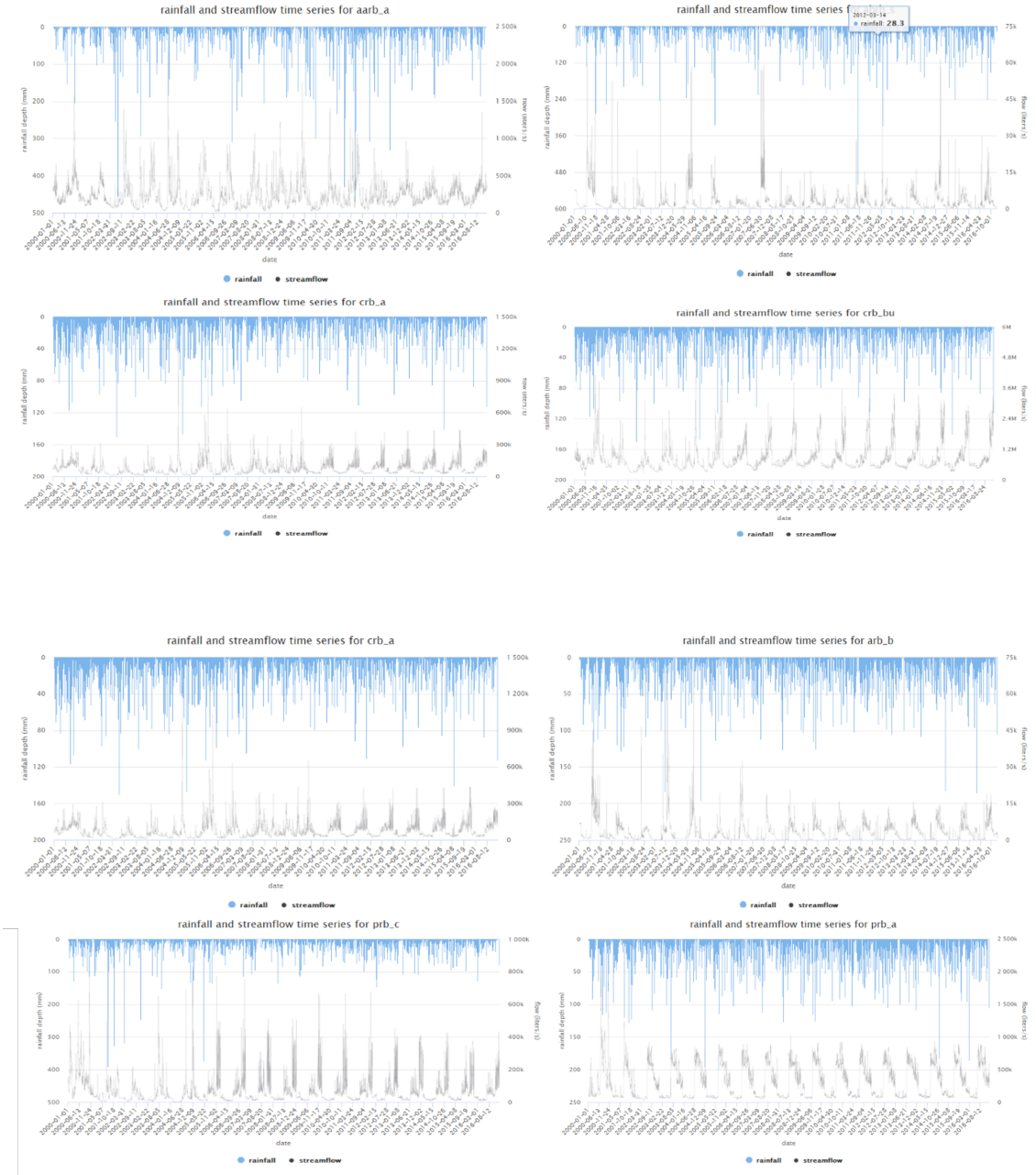



Partial Dependence on "LC9.CN"
Partial Dependence on "LC11.fl"
Partial Dependence on "LC10.man"
Partial Dependence on "LC2.for"

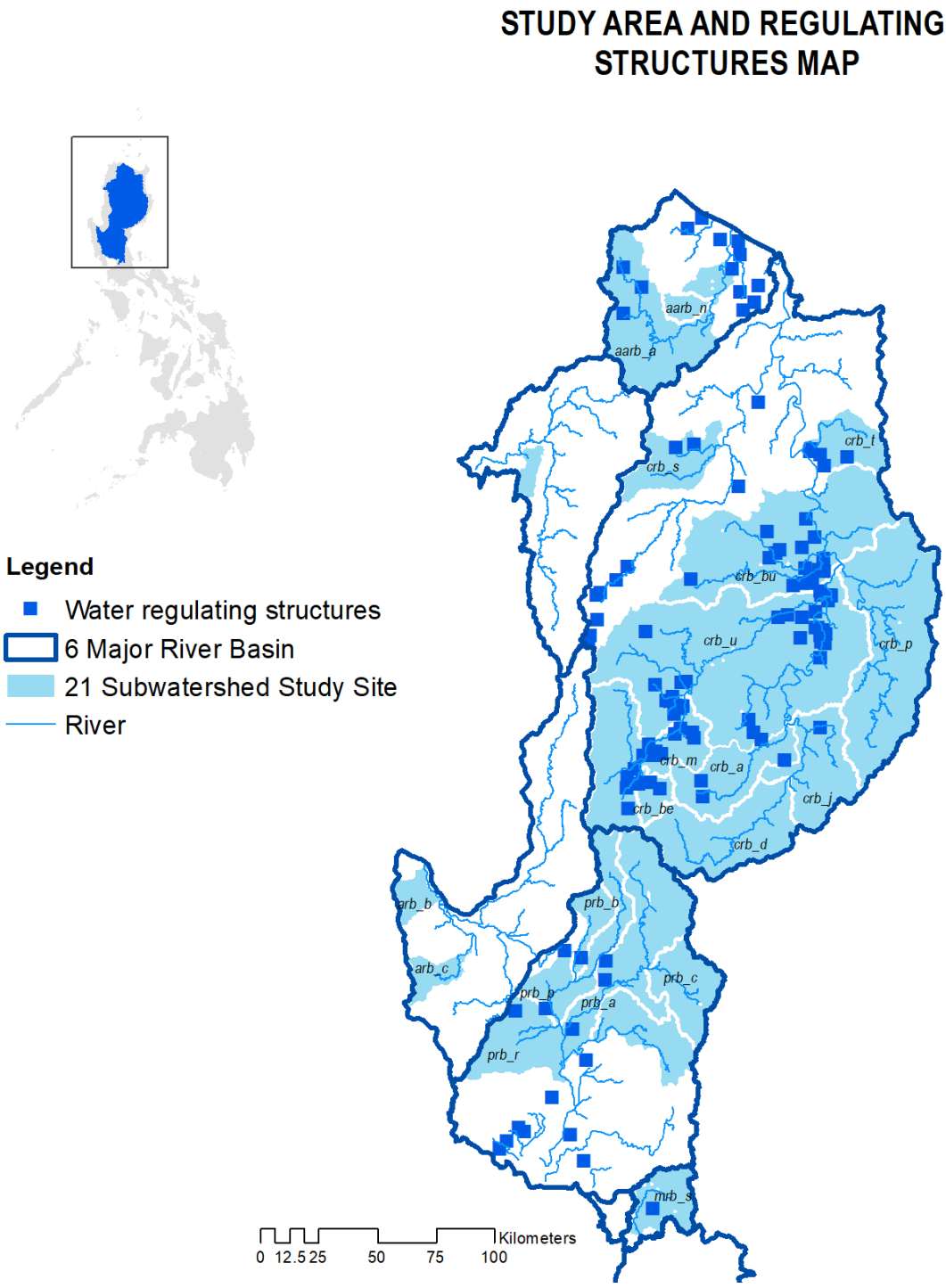
G Precipitation and predicted wet season streamflow correlation graphs.

Subwatershed	Accuracy (R2)	Climate type	Size	Number of water regulating structure	Rainfall-responsiveness	Deviation of historical flow	Correlation of forest loss and seasonal flow	Dry season streamflow		Wet season streamflow	
								starting year (2000)	*ending year (2016)*	*starting year (2000)*	*ending year (2016)*
aarb_a	0.56	3	large	5	yes	no data	no	156,837	146,021	402,897	370,126
aarb_n	0.72	3	small	0	yes	no data	yes	12,581	9,780	26,156	43,596
abrb_s	0.4	1	small	2	yes	no data	yes	452	447	5,752	6,150
arb_b	0.82	1	small	0	yes	yes	no	1,294	274	4,686	8,507
arb_c	0.1	1	small	3	no	no data	no	3,820	6,599	29,922	52,224
crb_a	0.62	4	large	2	yes	no data	yes	71,243	39,681	147,271	174,670
crb_be	0.21	3	small	3	no	no data	yes	1,622	422	10,997	6,432
crb_bu	0.65	3, 4	large	15	yes	no data	yes	459,034	432,249	1,908,951	1,974,319
crb_d	0.4	4	large	0	yes	yes	yes	47,329	5,866	44,749	110,130
crb_j	0.62	4	large	0	yes	no data	yes	175,337	171,429	237,715	201,830
crb_m	0.51	3	large	15	no	yes	yes	147,038	42,340	173,240	140,005
crb_p	0.65	4	large	0	yes	no data	yes	481,055	198,976	1,340,786	1,375,914
crb_s	0.2	3	small	2	no	no data	yes	93,999	21,220	80,283	53,910
crb_t	0.1	4	small	5	no	no data	no	25,583	22,098	194,583	192,109
crb_u	0.18	4	large	15	no	no data	yes	378,145	342,393	1,013,418	1,556,532
mrb_s	0.65	1	small	1	yes	no data	no	7,744	8,144	95,882	67,339
prb_a	0.2	1	large	15	no	no data	no	1,136	4,205	20,238	36,450
prb_b	0.76	1	small	0	yes	no data	yes	98,056	10,434	28,403	157,843
prb_c	0.86	1,3	small	0	yes	no data	yes	122,029	99,116	591,571	606,927
prb_p	0.4	1	small	5	no	no data	yes	65,535	49,558	245,972	209,048
prb_r	0.91	1	large	1	yes	no data	yes	30,408	15,829	91,060	92,897

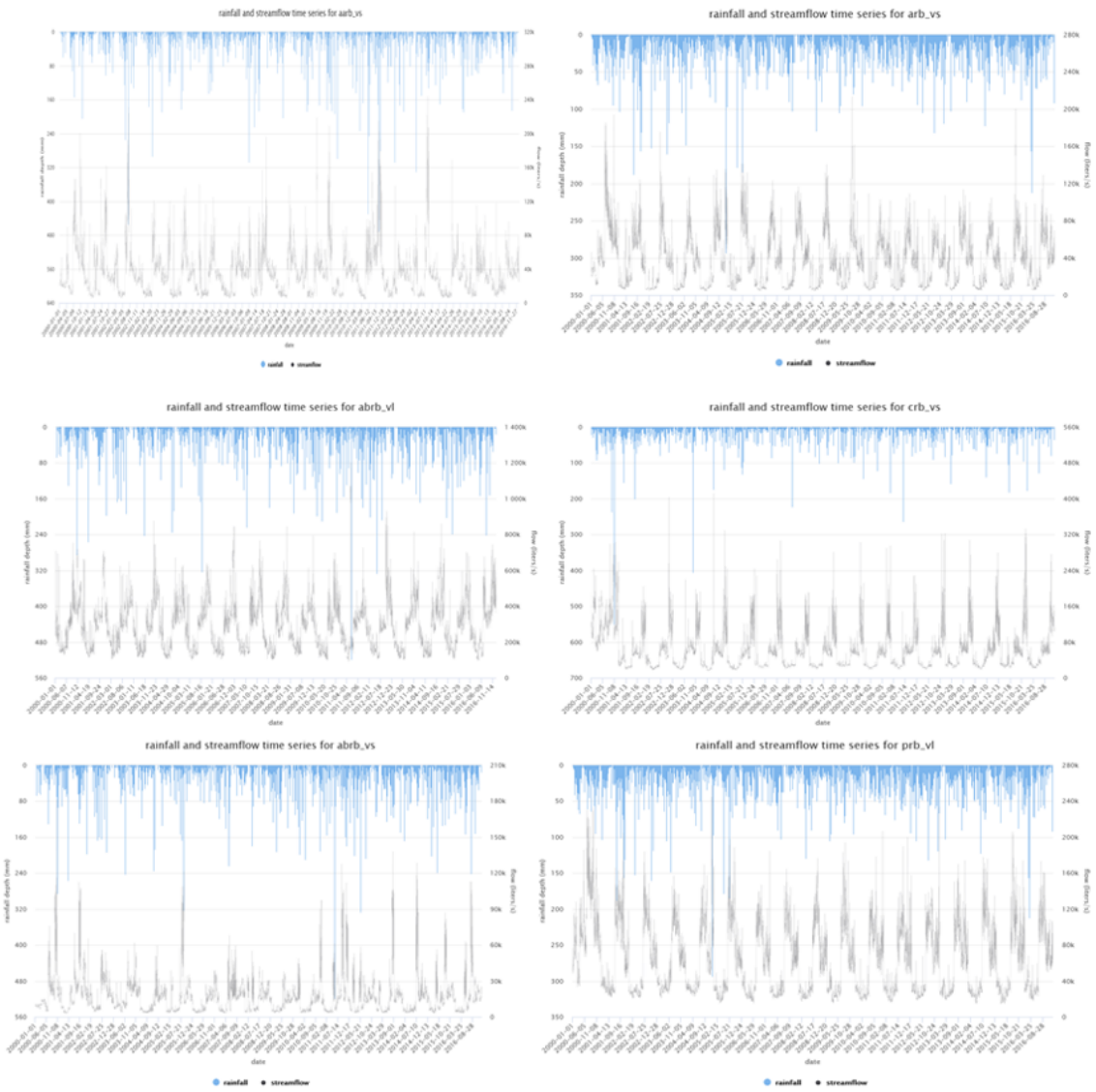
H Hydrographs for rainfall-responsive subwatersheds.



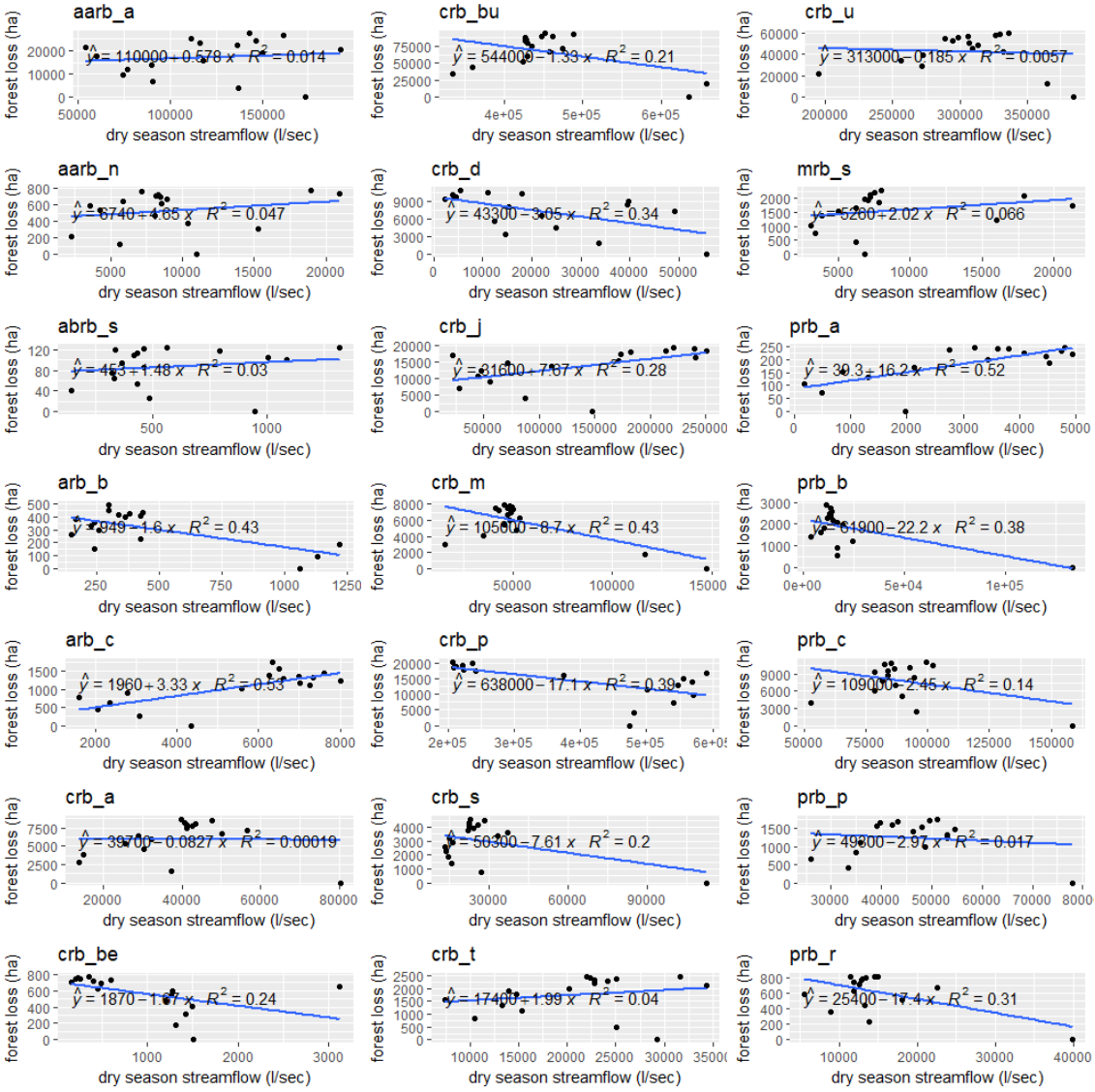
I Location of water regulating structures within the study area (source: basin master plans, RBCO.



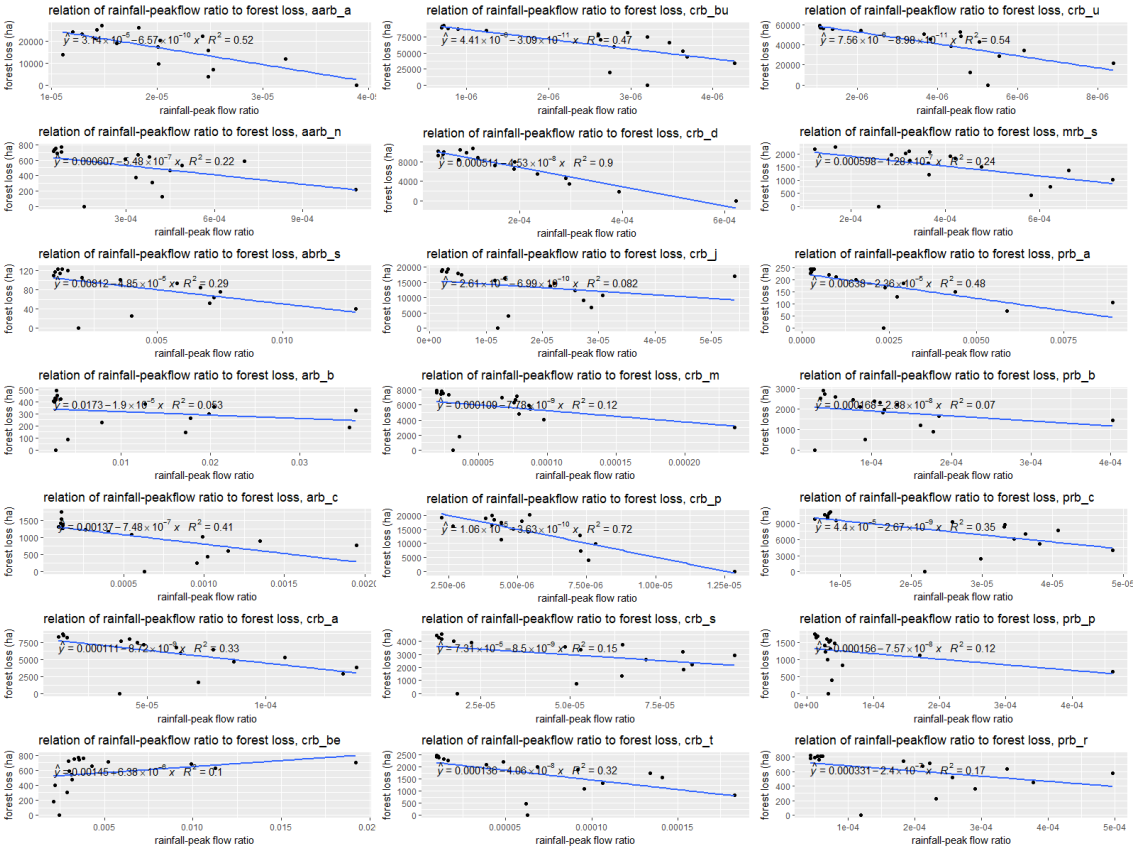
J Hydrographs for rainfall-responsive validation subwatersheds.



K Forest loss and predicted dry season streamflow correlation graphs.



L Peak flow-rainfall correlated with forest loss.



M Subwatershed area, forest area, and forest areas from 3 forest loss rate for dry and wet season flows.

Basin	% decrease in streamflow (dry season) relative to forest loss (fl) rate					
	fl rate 1	*fl rate 2*	*fl rate 3*	*% decrease, fl rate 1*	*% decrease, fl rate 2*	*% decrease, fl rate 3*
arb_b	0.04	0.09	0.14	0.53	0.68	0.82
prb_r	0.02	0.07	0.12	0.31	0.36	0.40
crb_m	0.01	0.06	0.11	0.16	0.24	0.32
crb_d	0.02	0.07	0.12	0.19	0.32	0.46
aarb_vs	0.02	0.07	0.12	0.68	0.72	0.76
abrb_vs	0.01	0.06	0.11	0.30	0.31	0.33
prb_vl	0.02	0.07	0.12	0.22	0.25	0.28
Basin	**Forest area (ha), forest loss (fl) rates in %**					
	total area (ha)	*forest area (2016)*	*% area (forests)*	*forest area, fl rate 1*	*forest area, fl rate 1*	*forest area, fl rate 1*
arb_b	31,609	1,981	6%	1,965	1,754	1,420
prb_r	147,804	7,159	5%	7,147	6,628	5,711
crb_m	215,601	71,897	33%	71,850	67,380	58,919
crb_d	169,590	100,788	59%	100,595	93,012	79,854
abrb_vs	19,927	4,766	24%	4,570	4,308	3,808
aarb_vs	34,625	25,965	75%	25,946	24,175	21,043
prb_vl	99,022	60,354	61%	60,297	58,526	55,393

Basin	% increase in peak flow-rainfall ratio (wet season) relative to forest loss (fl) rate					
	fl rate 1	*fl rate 2*	*fl rate 3*	*% decrease, fl rate 1*	*% decrease, fl rate 2*	*% decrease, fl rate 3*
aarb_n	0.02	0.07	0.12	0.93	0.94	0.95
crb_d	0.02	0.07	0.12	0.83	1.05	1.28
crb_p	0.03	0.08	0.13	0.88	1.10	1.33
Basin	**Forest area (ha), forest loss (fl) rates in %**					
	total area (ha)	*forest area (2016)*	*% area (forests)*	*forest area, fl rate 1*	*forest area, fl rate 1*	*forest area, fl rate 1*
aarb_n	19,277	18,124	94%	18,113	16,885	14,712
crb_d	169,590	100,788	59%	100,595	93,012	79,854
crb_p	1,187,380	249,510	21%	249,242	229,724	197,195