# Data-driven streamflow modelling in ungauged basins: to regionalize or not?

Arnan Araza[1,2], Lars Hein[1], and Confidence Duku[1]

## INTRODUCTION

- Streamflow models are crucial for large-scale water resources management but streamflow data is scarce especially in the tropics.
- Predicting streamflow in ungauged basins (PUB) can be categorized into: hydrological model-based (HMB) or data-driven streamflow models (DDS).
- The latter is gaining attention in PUB domain due to machine learning (ML) as a complementing model to HMB or as a stand-alone DDS often in local-scales.
- In large-scales i.e. regional modelling in tropics, use of ML is hindered by regional variability and streamflow data scarcity tandem.
- This can be dealt by "regionalization" which transfers model or model parameters of gauged to ungauged watersheds in a donor-receiver logic.
- We did a two-step approach to create a hybrid regionalized DDS model using Random Forest (RF) and compared it to a non-regionalized and other similar methods.
- We implemented this study in Luzon, Philippines – the country's largest island (Figure 1).

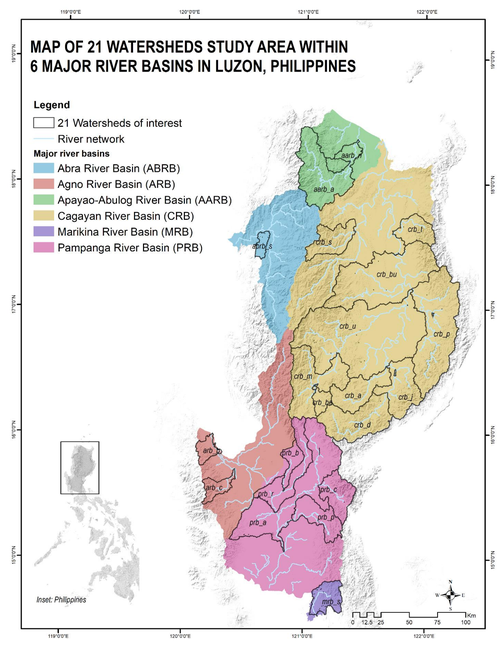


*Figure 1. Study area of 21 watersheds within the 6 major river basins in Luzon where the top rice-producing but flood-prone areas are.*

## OBJECTIVES

1. Compare different streamflow models and assess the ideal model for PUB regional modelling
2. Explain how regional variability affects streamflow models
3. Assess the error and uncertainty sources of the streamflow models

## METHODOLOGY

Using 51,690 streamflow data and 56 covariates (all spatial-open data) to characterize watersheds physically and climatically, four RF-based (Figure 2) DDS models were trained, compared, and assessed. Models 1-3 allowed watershed information grouping. See Table 1 for details.

*Table 1. Four RF-based DDS models showing the number of RF models and modelling details.*

| Model name | RFs | Details |
|---|---|---|
| Regionalized | 4 | Watersheds clustered into 4 based on Principal Components (PC) |
| Non-regularized | 1 | Lumped grouping where all watershed information combined |
| Semi-regularized | 6 | Watersheds grouped according to mother river basin |
| Localized | 21 | Each watershed has a stand-alone RF model |

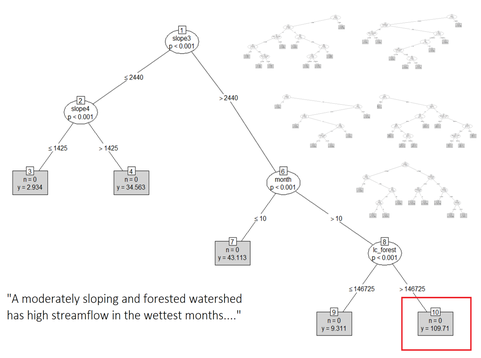


*Figure 2. Growing trees from randomized dataset which randomly selects root node (top circle) then splits the succeeding nodes by assuring minimal variability and stops until prediction is made.*

Predicted streamflow from all models were validated with independent data. We then assessed variability, uncertainty, and bias as per RF model and covariates. The importance of covariates were based on variable importance metric or VIM (sensitivity to permutation or shuffling values).

## CONCLUSIONS

- Regionalized DDS is ideal for PUB regional modelling as reflected by the higher accuracy, reduced uncertainty and lower bias.
- Dealing with regional variability prior to model training is strategic.
- Static variables becomes more useful after regionalization

## MAIN RESULTS AND DISCUSSION

**Main result 1:** From the validation (Figure 3), regionalized model was the most accurate by at least 25% from other models and with lesser prediction uncertainty than semi-regionalized and non-regionalized models.

**Discussion:** The PC-based clustering explained 80% of the dataset variability. Without it, complex mixed information from watersheds can cause model inaccuracy in addition to limited training data. This is also reflected with the prediction uncertainty where the localized model had the least by being ungrouped or no regional variability.

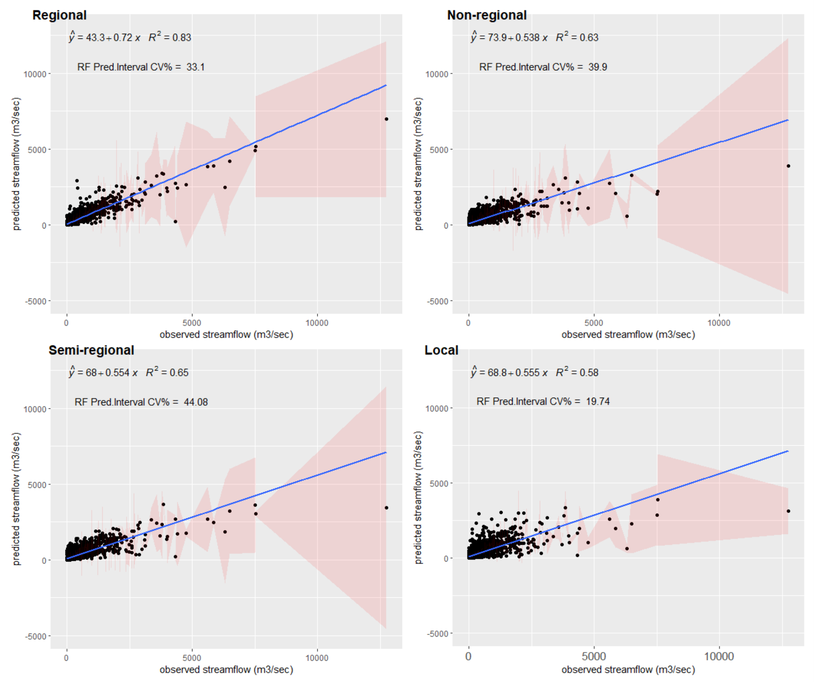


*Figure 3. Validation results using all test data of the 4 RF models With $R^2$, Coefficient of Variation %, and model prediction interval (red).*

**Main result 2:** When regionalizing, static covariates became useful (0 to 22% VIM increase). Soil and weather covariates constituted 70% and 79% of the top important static and dynamic covariates, respectively. See Figure 4.

**Discussion:** Regionalization facilitates better information grouping based on physical covariates and streamflow. Use of high resolution soil data (ISRIC, 250m) allows better soil representation while weather data is daily and the most sensitive to permutation.

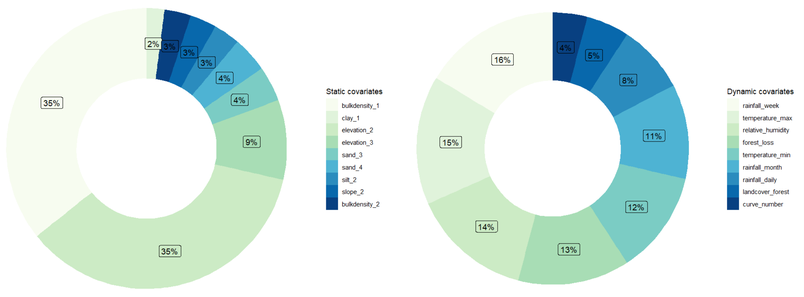


*Figure 4. Top static and dynamic covariates from relative VIM.*

**Main result 3**: Aggregation into monthly streamflow (Figure 5) highlighted the bias (mean error). There was over-prediction in most of the months except for the wettest months. Most unbiased was the regionalized model.

**Discussion:** Given that models are sensitive to weather covariates, more "rainy days learning" attributed to pseudo-rainy days (dams, post-typhoons, mixed information) can cause bias. There is also a systematic model prediction bias to the mean.

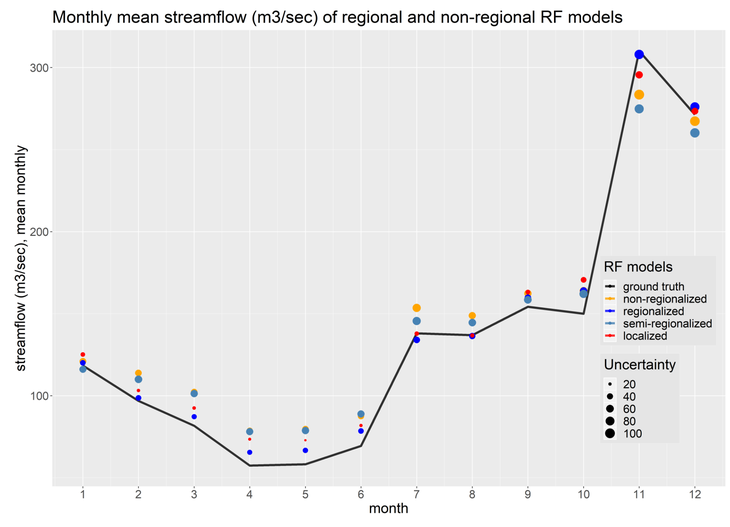


*Figure 5. Mean monthly streamflow of the 4 RF models to highlight bias and uncertainty relative to observed (ground-truth) data.*

[1] Environmental Systems Analysis (ESA) and [2] Geo-Information and Remote Sensing (GRS), Wageningen-UR, Wageningen, The Netherlands; [1,2] email address: arnan.araza@wur.nl