Global monitoring of fresh water at high spatial and temporal resolution.

Assessing stream and lakes hydrological/physical features whitin a machine learning framework

PI: Dr. ***Giuseppe Amatulli* -** Research Scientist at F&ES and YCRC, giuseppe.amatulli@yale.edu, Tel: 6057289031
Expertise: Geocomputation, data analysis, GIS, remote sensing , big geo-dataset processing, hydrology.

The overarching goal of this research will be to revolutionize our understanding of the fundamental principles that govern water regimes in streams and lakes worldwide. The project will capture the multi-dimensional aspects of the flow regimes (river discharge) and model hydraulics using a wide range of high resolution geo-datasets, while gauging station data (observations) in a machine learning framework. A challenging project is described herein.

Background: Freshwater is among the most vulnerable resources in nature and is vital for all organisms. Lakes and rivers contain a significant quantity of this resource, and provide the basis for water discharge quantification and modelling of the hydrological system process. Freshwater quantification at high resolution/accuracy is the first step for a global comprehensive of the overall water cycle. Current knowledge of stream flow trends in developing countries and/or in rural areas is very limited and fragmented, especially since small streams are not represented. A full geo-analysis is needed to capture these stream features, which can subsequently be shared in the public domain. By understanding and quantifying existing hydrological patterns, it is possible to reduce flooding risks[1], estimate sediment transportation[2], quantify hydropower production capacity[3] and drinking water supply[4,5], determine the role of inland waters in greenhouse gas budgets[6,7], and preserve freshwater biodiversity[5,8]. These findings will also provide guidelines at a global-scale for conservation strategies, water management, flood prediction and climate change mitigation.

The physical characteristics of each watershed or stream are the result of complex interactions among several environmental variables that regulate the discharge regimes and stream profiles, such as rainfall, evapotranspiration, soil infiltration and retention, geomorphology, land use, and snow cover, among others. The upstream environment influences the physical characteristics of a stream in any given location, and adds complexity to model the system. The research will conduct a comprehensive quantification of hydrological features of these water bodies, offering ground-breaking analyses of flow regimes within spatial and temporal domains.

Currently, the available global hydrographies are derived from Digital Elevation Models (DEMs) and released at a coarse spatial resolution (HydroSHEDS[9] 500m, Fig.1). Furthermore, small streams are usualy underestimated[10]. Another limitation is the fact that the streamflow data (monthly discharge in m^3/s) recently derived from HydroSHEDS (FLO1K[11]) lack an explicit assessment of fine geographic/temporal variations and detailed stream hydraulic features.

From a methodology point of view, the prediction of flow regimes has been based primarily on either empirical or physically-based models. The latter simulate a simplified watershed system and express stream behavior by solving mathematical equations that reflect hydrological processes. These models provide accurate estimates of flow for ungauged watersheds, but the model parameters are typically difficult to estimate. Hence, these models still require *ad-hoc* calibration unique to each watershed, which makes “up-scaling” to a global level rather arduous. Such models are also criticized for being over parameterized[12]. The general understanding of this multifaceted system is still fragmented[11] and predominantly studied at catchment scale.

Objectives: In this project, we will implement a data-driven approach, using several geo-datasets that control the flow regimes addressing a suite of long-standing and open-ended questions, such as: (i) What are the multidimensional relationships among environmental variables that influence stream characteristics? (ii) How can we measure long-term (past 30 years) monthly changes of water discharge at high resolution (90 m) globally? (iii) How will we predict water regimes in small and ephemeral streams? (iv) Are remote sensed water occurrence products useful for estimating the time of concentration in small and large rivers?

To address these questions, we will use a novel experimental approach to derive a global stream network from high resolution DEM (90m), as opposed to the previous products (HydroSHEDS[9] 500m), and build a suite of freshwater-specific seamless environmental variables (90 m resolution) that will integrate the upstream environmental conditions[14]. This procedure will generate a set of covariates, from which spatially-explicit machine learning algorithms will derive accurate stream hydraulic characteristics worldwide, thereby improving FLO1K[11].

Geocomputational Implementation: The overall Geocomputational process will be implemented using the High Performance Computing (HPC) clusters of the Yale Center for Research Computing. Open-source geographical software and complex scripting routines will be used to run the full computation. The complete workflow can be summarized in three main phases:

1. **Derive stream networks from DEM**: Substantial improvements in computational power, advanced scripting capability, and a high resolution DEM, such as MERIT[15], will allow the accurate calculation of water drainage using flow accumulation algorithms. The MERIT-derived stream network (Fig. 1) will yield the trajectory of stream channels and the relative catchments and sub-catchments areas of each stream segment worldwide.
2. **Calculate freshwater environmental variables:** Environmental information will be harmonized and attributed to each sub-catchment (Fig. 2) area using a hierarchical-nesting approach[14], which will trace the flow connectivity along the stream network[8]. Missing data among the environmental variables will be estimated using nonparametric regression techniques (e.g. kernel smoothing and interpolating splines) considering the variation on the temporal/spatial domain. High resolution variables will be selected to match the time span of the Global Surface Water[16] occurrence product so as to allow a prediction and validation across the past 30 years.
3. **Model stream features:** The availability of harmonized databases of observations at worldwide hydrological gauging stations[18], in combination with the precompiled freshwater variables, will serve as the foundation for implementing machine learning techniques capable of deriving stream characteristics. Several approaches for regression tasks will be tested, such as *artificial neural networks*, *random forests*, and *gradient boosting*. Resampling techniques such as the jackknife will be employed to estimate confidence intervals[17,18]. These techniques have demonstrated strong predictive capabilities, particularly when applied to large high-dimensional datasets. Some of these techniques have been applied to the estimation of annual streamflow with great success, but only at coarser resolutions[11] or at catchment scale[19]. The models we obtain will be useful to predict monthly minimum, mean and maximum water regimes on ungauged streams and lakes around the world in the last 30 years.

Thecnical aspects: We estimate handling 100 TB of geo-data by using a complex scripting procedure which will use different programming languages (Bash, Python, R, Julia, AWK, GRASS, GDAL, CDO, PKtools). The proposed research is unique not only from a computational perspective (volume of data), but also due to the use of advanced machine learning techniques at a previously unachieved global-scale/fine-resolution. Given the novelty of the research and its broad-scale implementation, we will apply for an external grant in 2019, enlarging the horizon for massive geo-data processing at Yale. The research will be published in high-impact peer-reviewed publications. In the spirit of reproducible research, we will also release the scripting procedure. The team will rely on individuals with strong scientific/statistical backgrounds and proficient programming skills to develop the workflow. Dr. Amatulli will coordinate the project and actively work with the Post-Doc to delineate the overall scripting procedure. The postdoc will be selected on the basis of his/her computational/programming skills, geographic/hydrological background, statistical computing and machine learning experience. The funds will be used to cover the two-year salary for a postdoc position which will ennance the modelling framework and two-year salary for a GIS-WEB deloper wich will build up the web platform for data visualization and data sharing.

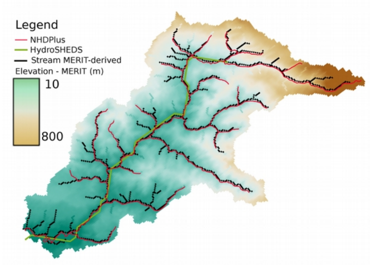

Figure 1: MERIT-DEM stream network delineation and comparison with other available stream networks.

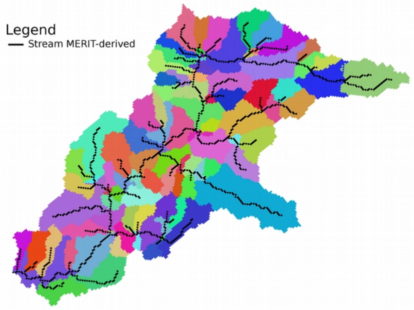

Figure 2*:* Sub-catchment delineation, based on the stream network derived from MERIT-DEM.

References [1] Winsemius, H. C., et al., *Nature Climate Change*, 6, 4, 2016; [2] Hilton, R. G., et al., *Nature Geoscience*, 1, 2008; [3] Van Vliet, M.T.H. et al., *Nature Climate Change*, 6.4 , 2016; [4] Michalak, A. M., *Nature* 535, 2016; [5] Vörösmarty, C. J. et al., *Nature*, 467, 2010; [6] Raymond, P. A. et al., *Nature*, 503, 2013; [7] Butman, D., et al., *Nature Geoscience*, 4, 2001; [8] Domisch, S., et al., *Ecography*, 39, 2016; [9] Lehner, B., , et al. Eos, Transactions American Geophysical Union 89, 2008; [10] Benstead, J.P., et al., *Nature Geoscience* 5, 10, 2012; [11] Barbarossa V. et al., Scientific Data, 2018; [12] Yamazaki, D., et al., *Nature* 540, 7633, 2016; [13] Kuppel, Sylvain, et al., *Environmental Modelling & Software*, 101, 2018; [14] Domisch, S., et al., *Scientific Data*, 2, 2015; [15] Yamazaki, Dai, et al., *Geophysical Research Letters*, 2017; [16] Pekel, J.F., et al., *Nature*, 540, 2016; [17] Efron, B. *Breakthroughs in statistics*. 1992 [18] Wager, S, et al. *The Journal of Mach. Learn. Res.* 15.1, 2014; [19] Worland, SC, et al. *Environmenu. Modell. & Soft.* 101, 2018.