Data-based machine learning unveils ecosystem metabolic regimes at the scale of entire stream networks

Pier Luigi Segatto¹, Tom J. Battin¹, and Enrico Bertuzzo²
¹École polytechnique fédérale de Lausanne, Environmental Engineering Institute - IIE, SCHOOL OF ARCHITECTURE, CIVIL AND ENVIRONMENTAL ENGINEERING-ENAC, Lausanne, Switzerland (pier.segatto@epfl.ch)
²Department of Environmental Sciences, Informatics and Statistics, University of Venice Ca' Foscari, 30170 Venice, Italy

Inland waters are major contributors to the global carbon cycle. Nowadays, new sensor technology has changed the way we study ecosystem metabolism in streams. We are able to produce long-term time series of gross primary production (GPP) and ecosystem respiration (ER) to infer drivers of the stream ecosystem metabolic regime and its seasonal timing. Despite big data availability, most studies are limited to individual stream reaches and do not allow the appreciation of metabolic regimes at the scale of entire networks, which, however, would be fundamental to properly assess the relevance of metabolic fluxes within streams and rivers for carbon cycling at the regional and global scale. Machine learning (ML) has great potential in this direction. Firstly, ML could be used to extrapolate both in time and space heterogeneous forcings (e.g., streamwater temperature (T) and photosynthetic active radiation (PAR)) required to run a process-based model for reach-scale metabolism to the scale of an entire stream network. Secondly, the same procedure could be applied to reach-scale estimates of ecosystem metabolism to check whether available data contain enough information to explain the network scale variability. In this study, we used Random Forest to predict patterns of environmental forcings (T and PAR) and stream metabolism (GPP and ER) at the scale of an entire stream network. We used available high-frequency measurements of T and PAR, estimates of ecosystem metabolism and major proximal controls (e.g., incident light, discharge, stream-bed slope, drainage area, water level, air temperature) from twelve reaches within the Ybbs River network (Austria) and explicitly trained our Random Forests by integrating distal factors, namely: vegetation type, canopy cover, hydro-geomorphic properties, light, precipitation, and other climatic variables. We designed two different training setups to assess spatial and temporal predicting model capabilities, respectively. This approach allowed us to reliably infer the target variables (T, PAR, GPP, and ER) on annual basis across a stream network, to filter the most important predictors, to assess the relative contribution of the metabolic fluxes from small to large streams, to estimate annual metabolic budgets at different spatial scales and to provide empirical evidence for long-standing theory predicting shifts of ecosystem metabolism along the stream continuum. Finally, we estimated autochthonous and allochthonous respiration for the entire stream network, which is crucial to integrate the role of ecosystem processes for the carbon cycle.