EMPLOYING UNSUPERVISED MACHINE LEARNING ALGORITHM TO IDENTIFY CRITICAL SOURCE AREAS TO ENHANCE WATER QUALITY: INTEGRATING SWAT MODELING AND MULTI-VARIABLE STATISTICAL ANALYSIS

***Shubo Fang**, Matthew J. Deitch, Tesfay G. Gebremicael*

Soil, Water, and Ecosystem Sciences Department, University of Florida/IFAS/West Florida Research and Education Center, Milton, FL, USA

Non-point source pollution (NPS) management remains a pressing global challenge. Due to the characteristics of source diffusion, the random occurrence of spatiotemporal patterns, and the multiple underlying processes that directly or indirectly impact NPS, effective management strategies for NPS remain difficult to initiate and implement. The concept of critical source areas, or priority management areas (hereafter CSAs for brevity), has gained significant attention as an approach to efficiently identify and manage the primary sources of NPS. CSAs are areas that contribute disproportionately to NPS pollution and require targeted management interventions. Various measures have been proposed to identify CSAs. These measures can be categorized into two classes: those based on multi-variable statistical analyses and those based on physical-process modeling, such as the Soil and Water Assessment Tool (SWAT). By integrating SWAT modeling and land use and land cover (LULC) based multi-variable statistical analysis, this study aimed to identify driving factors, potential thresholds, and CSAs to enhance water quality in southern Alabama and northwest Florida's Choctawhatchee Watershed. An unsupervised machine learning algorithm, i.e., the self-organizing maps (SOM), also known as a Kohonen map or a Kohonen network, was employed for data visualization and clustering. The results revealed the significance of forest cover and of the lumped developed areas and cultivated crops ("Source Areas") in influencing water quality. The stepwise linear regression analysis based on SOMs showed that a negative correlation between forest percent cover and total nitrogen (TN), organic nitrogen (ORGN), and organic phosphorus (ORGP), highlighting the importance of forests in reducing nutrient loads. Conversely, Source Area percentage was positively correlated with total phosphorus (TP) loads, indicating the influence of human activities on TP levels. The receiver operating characteristic (ROC) curve analysis determined thresholds for forest percentage and Source Area percentage as 37.47% and 20.26%, respectively. These thresholds serve as important reference points for identifying CSAs. Based on the threshold of forest percentage (< 37.47%), it was determined that 46% of the entire Choctawhatchee Watershed fell within the identified CSAs. These areas accounted for approximately 67% of the total TN loads in the watershed. Similarly, applying the threshold of Source percentage (> 20.26%), it was found that 33% of the entire watershed was prioritized as CSAs, covering approximately 54% of the TP loads. But if considering both thresholds (forest percentage < 37.47% and Source percentage > 20.26%), it was identified that 28% of the entire watershed was prioritized as CSAs. These areas covered approximately 47% of the total TN loads and 50% of the total TP loads in the watershed. The study underscores the importance of considering both physical process-based modeling and LULC for a comprehensive understanding of watershed water quality management.

PRESENTER BIO: Shubo Fang is now a postdoctoral research associate at Soil, Water, and Ecosystem Sciences Department, University of Florida/IFAS/West Florida Research and Education Center, Milton. Dr Fang's research encompasses coastal wetlands degradation, coastal wetlands conservation and watershed ecology. By different kinds of quantitative methods Dr Fang try to reveal the interactions between the geo-surface pattern and the associated processes.