Title: Identification of Environmental Factors Controlling Biogenic Secondary Organic Aerosol Formation in the Central Amazon: a machine learning approach

Maria Eduarda Guedes Leal[1,2], Joel F. de Brito[1], Anna Font[1], Marcio Cataldi[2], Pedro Dezan[3]

1 - IMT Nord Europe, Douai, France

2 - Climate System Monitoring and Modeling Laboratory (LAMMOC), Federal Fluminense University , 156 Passo da Pátria St., 24210-240 - Niterói, Brazil

3 - Environmental Engineering Program, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro 21941-909, RJ, Brazil

Abstract: Biogenic secondary organic aerosols dominate submicron aerosol mass in the central Amazon during the wet season, yet their molecular composition and formation pathways remain difficult to resolve due to the complexity of precursor chemistry and atmospheric processing. Recent measurements using a chemical ionisation mass spectrometer (PTR-Qi-ToF-MS-CHARON) in Central Amazon, at the Atmospheric Tall Tower Observatory (ATTO) during the wet season of 2022 provided near–molecular-level characterization of SOA tracers associated with isoprene-, monoterpene-, and sesquiterpene-derived oxidation pathways. Previous analysis of these observations reconstructed total organic aerosol (OA) concentrations using Multiple Linear Regression based on selected molecular tracers, explaining approximately 75% of OA variability and suggesting that non-linear chemical interactions may contribute to the remaining unexplained fraction. In this study, Random Forest algorithm is applied as an interpretative machine learning approach to improve OA reconstruction and investigate their relationships with molecular tracers. The analysis focuses on the same tracer set identified in chamber and field experiments and evaluates model behavior across four contrasting atmospheric regimes observed during the campaign: pristine conditions, long-range transport influence, high OA loading, and post-rainfall recovery. The analysis will assess whether Random Forest improves the explanatory power of OA reconstruction relative to Multiple Linear Regression and explore potential non-linear relationships between molecular tracers and OA concentrations across the atmospheric regimes observed during the campaign. Feature importance analysis will provide insights into the relative contribution of individual tracers and their combined influence on SOA variability. By integrating machine learning with high-resolution molecular measurements, this work aims to provide new insights into SOA formation processes in the Amazon and demonstrate the potential of interpretable artificial intelligence methods for atmospheric chemistry and climate-relevant aerosol research.

Keywords: Atmospheric Chemistry, Climate Change Adaptation, Machine Learning Interpretability, Non-linear Climate Interactions, Random Forest, Secondary Organic Aerosol