

Improved Gap-Filling of Eddy Covariance Fluxes Using Remote Sensing and Environmental Variables via XGBoost

Simon De Cannière¹, Sebastian Wieneke², Thomas Servotte³, Adrià Descals¹, Tim Verdonck³, Ivan Janssens¹

¹: Plants and Ecosystems (PLECO), Biology department, University of Antwerp

²:Institute for Earth System Science and Remote Sensing, Leipzig University

³:Mathematics department, University of Antwerp

Introduction

Eddy covariance requires (i) turbulence conditions and (ii) tight hardware requirements, causing major gaps in the data

The default ICOS gapfilling method deals with Marginal Distribution Sampling (MDS), in which an ad-hoc LUT is built based on data before and after the gap, but this approach is oblivious to vegetation dynamics or soil water dynamics

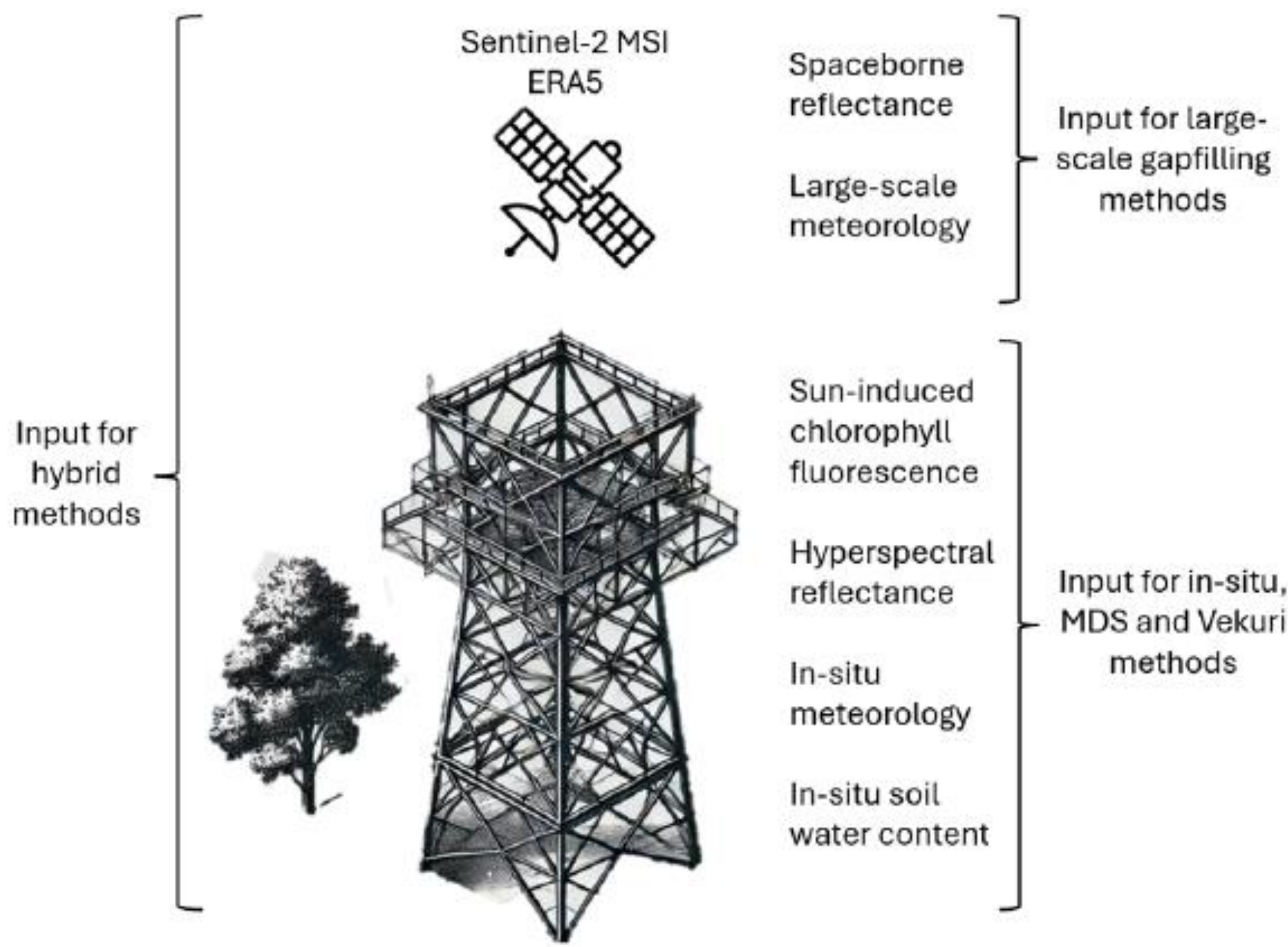
MDS has trouble following diurnal dynamics in carbon fluxes

Tree-based estimators like XGBoost can consider many input variables, including environmental drivers, data on vegetation and proxy variables on photosynthesis

Sun-Induced Chlorophyll Fluorescence (SIF) is a potent proxy for photosynthetic activity

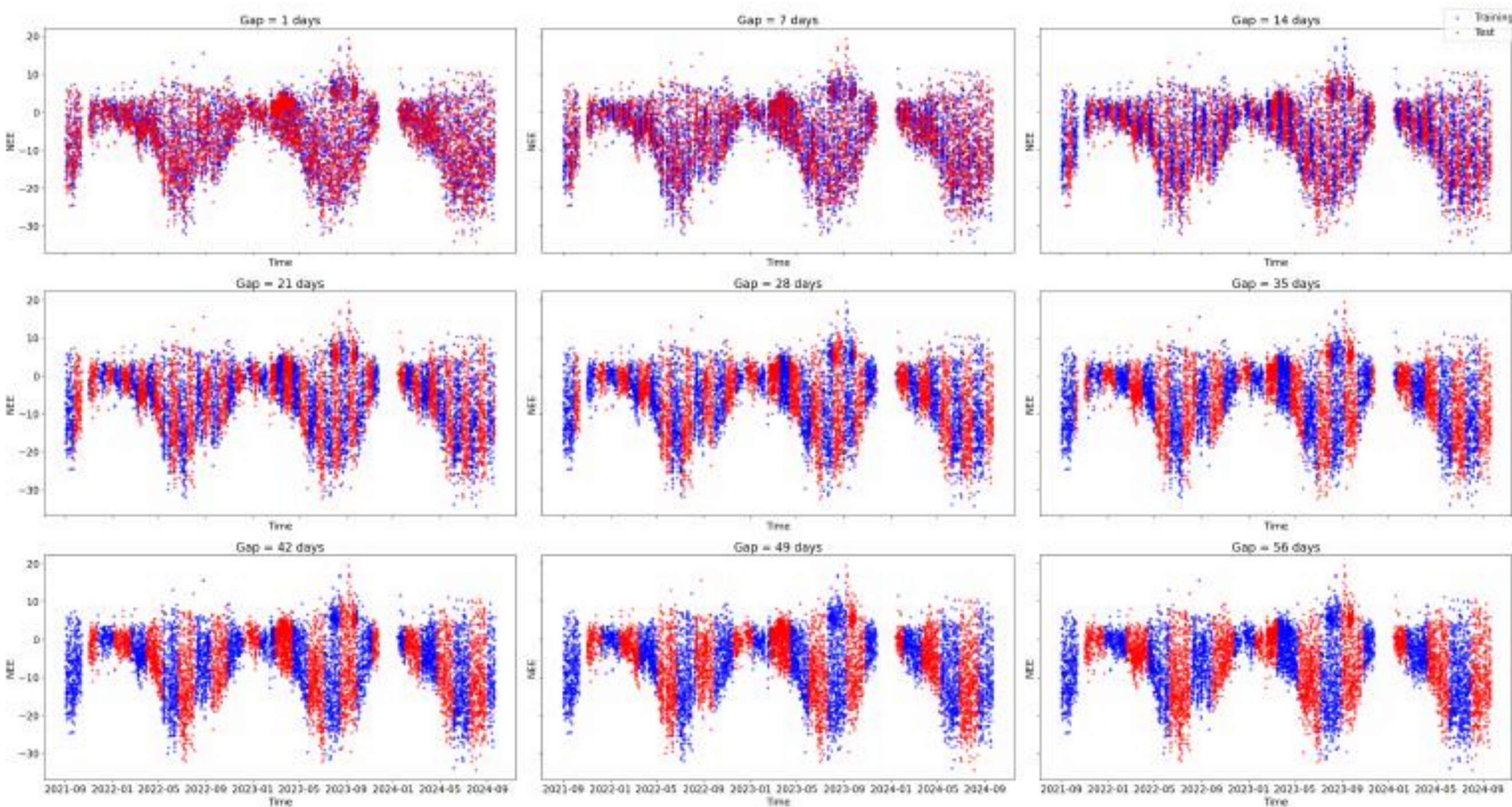
Setup

Four years of in-situ data of meteorology, soil water content and SIF at BE-Bra site, combined with land surface data from ERA5 and satellite-based data from Sentinel-2. With that four XGBoost model scenarios are defined: (i) one with in-situ data (in-situ) (ii) one with ERA5 and Sentinel data (large-scale), (iii) a hybrid approach of in-situ PAR and two benchmark scenarios using Air Temperature, Vapour Pressure Deficit (VPD) and irradiation as input. As algorithm either XGBoost (Vekuri approach) or MDS is used.



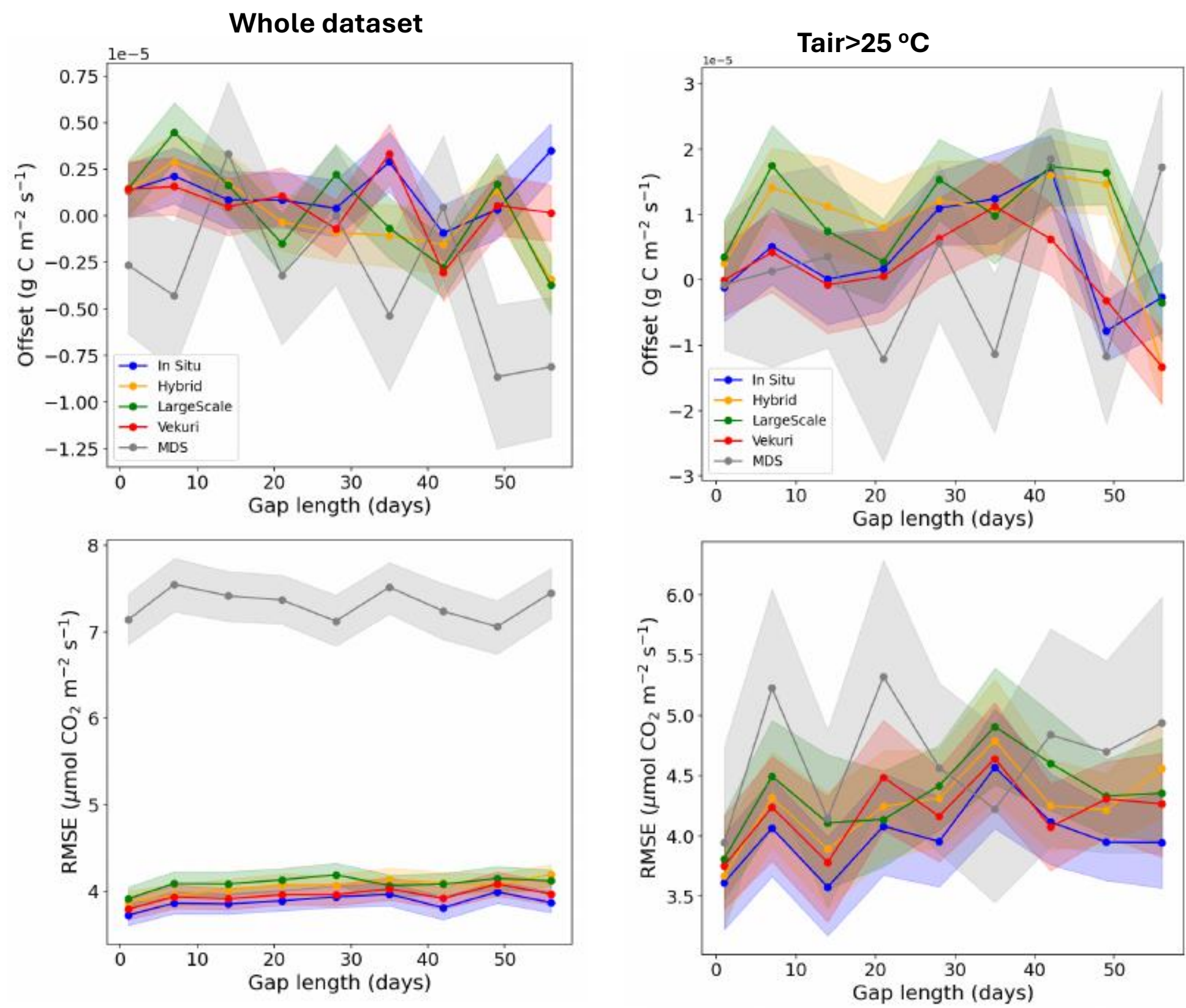
Splitting train and test data

The dataset was split into a 50% training and 50% test data using a ‘zebra’-splitting with increasingly long gaps, from 1 day to 49 days in length



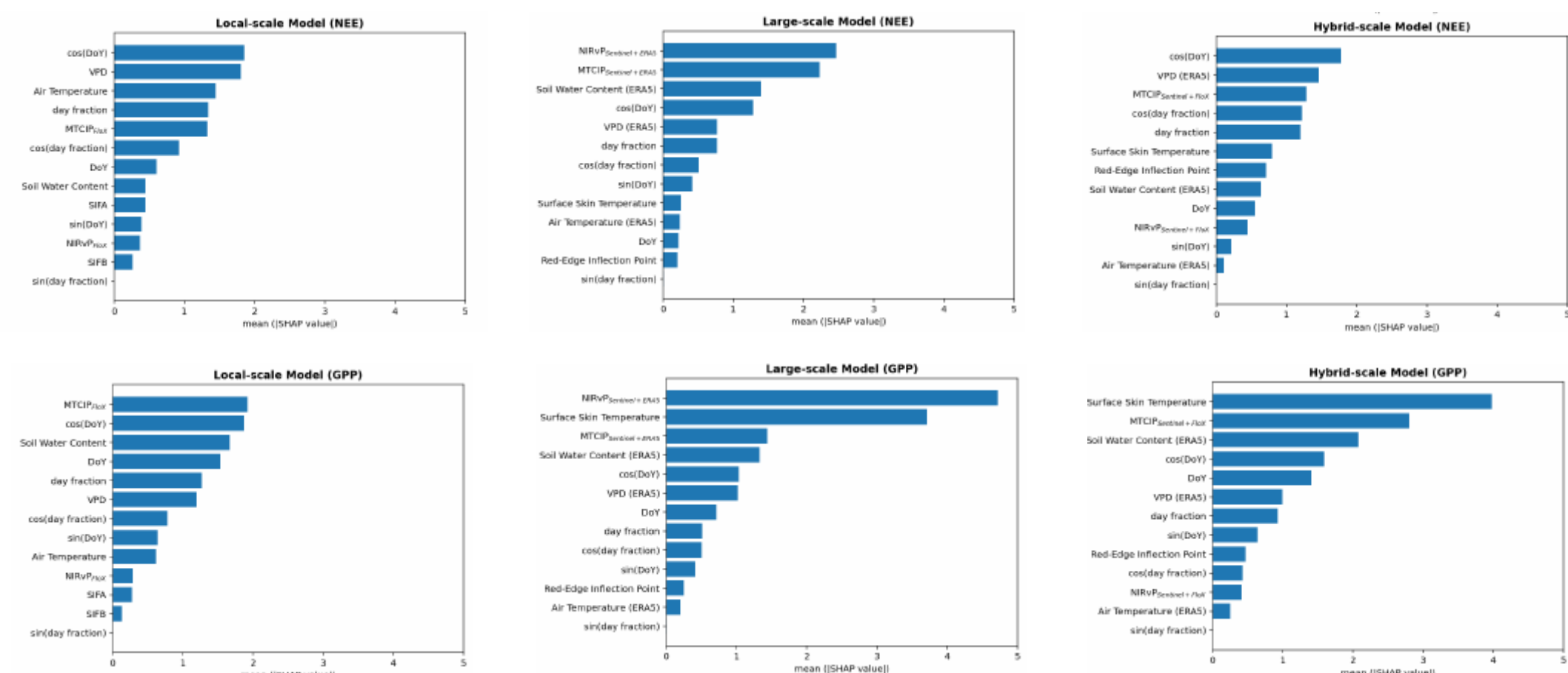
Performance of XGBoost models

XGBoost-based turned out to be more stable compared to the MDS-based model. RMSE was remarkably lower for XGBoost-based models compared to the MDS-based model. The in-situ model outperformed the other models when focussing on warm data



Explanatory power of input data

Classical drivers of light, air temperature and VPD turn out to be most important, in addition to time-related variables, most notably cos(DoY). Soil water content and SIF mainly were important in high-temperature conditions. For the large-scale GPP model, the Surface Skin Temperature contained important information, carrying an imprint of both the weather and of the stomatal dynamics.



Conclusion

XGBoost can consider many variables in its estimation for NEE and GPP. Adding soil moisture and sun-induced chlorophyll fluorescence mainly showed its value when only considering the warm datapoints. Data from land surface models or satellites showed their added value in improving the estimation of carbon fluxes.