EGU25-9368, updated on 12 Jan 2026
https://doi.org/10.5194/egusphere-egu25-9368
EGU General Assembly 2025

Assessing the optimal drivers for flux data gap-filling using random forest networks

Nicola Lieff[1,2], Daniel Metzen[3], Cacilia Ewenz[1,2,5], Peter Isaac[5], Ian McHugh[5,6], and Anne Griebel[4]
[1]Airborne Research Australia, Adelaide, Australia
[2]Faculty of Science Engineering and Technology, University of Adelaide, Adelaide, Australia
[3]Western Sydney University, Penrith, Australia
[4]School of Life Sciences, University of Technology Sydney, Sydney, Australia
[5]TERN Ecosystem Processes Central Node, James Cook University, Cairns, Australia
[6]School of Ecosystem and Forest Sciences, University of Melbourne, Richmond, Australia

The Terrestrial Ecosystem Research Network (TERN) OzFlux group operates a network of eddy covariance stations that collect long-term atmospheric and soil measurements for monitoring and understanding changes in climate and the environment. Ideally, all data collected would be gap-free, however, all real data has gaps where instruments have not recorded measurements or data has been discarded due to low turbulence. To allow this data to be used as a continuous time-series in further analysis, the missing data is gap-filled using PyFluxPro. The standard community approach uses a predefined set of variables (drivers) for gap-filling, which are the same variables for all stations irrespective of location. However, the stations are located in a large range of climate zones, hence the standard gap-filling drivers might not be ideal for all sites. This is because the drivers were chosen for a small set of initial sites and might not be representative for a heating and drying climate.

To identify which drivers were best suited for each station, we developed a random forest model to objectively assess the relative importance of input variables used to gap-fill ustar, carbon, and energy fluxes. We trained this model on the published TERN OzFlux data for all available Australian sites using a large range of input variables. This model then determined the relative importance of variables, mean absolute errors, and R^2 for the accuracy of the model prediction for a target variable at each site. Next, we grouped the variables into atmospheric, energy, turbulence and soil categories of drivers, which highlighted a distinct variation in the contribution of each category of driver across sites. To assess the ecological significance of these trends, the model importances were sorted by the aridity index and grouped by the Köppen-Geiger classification of each site. There is a notable shift in the importance of energy, turbulence, and soil groups with decreasing aridity, and driver contributions were generally consistent within Köppen-Geiger classifications. Reprocessing the gap-filling of a representative subsample of sites demonstrated a marked improvement in predicting the gap-filled target variables, highlighting that this approach can inform driver selection at new and established sites and will improve the understanding of the ecological significance of different drivers in various climate regions.