EGU2020-8823
https://doi.org/10.5194/egusphere-egu2020-8823
EGU General Assembly 2020


# Towards a generalized framework for missing value imputation of fragmented Earth observation data

**Verena Bessenbacher**, Lukas Gudmundsson, and Sonia I. Seneviratne
ETH Zürich, Institute for Atmospheric and Climate Sciences, Switzerland (verena.bessenbacher@env.ethz.ch)

The past decades have seen massive advances in generating Earth System observations. A plethora of instruments is, at any point in time, taking remote measurements of the Earth's surface aboard satellites. This birds-eye view of the land surface has become invaluable to the climate science and hydrology communities. However, the same variable is often observed by several platforms with contrasting results and satellite observations have non-trivial patterns of missing values. Consequently, mostly only one remote sensing product is used simultaneously. This and the inherent missingness of the datasets has led to a fragmentation of the observational record that limits the widespread use of remotely sensed land observations. We aim towards a generalized framework for gap-filling global, high-resolution remote sensing measurements relevant for the terrestrial water cycle, focusing on ESA microwave soil moisture, land surface temperature and GPM precipitation. To this end, we explore statistical imputation methods and benchmark them using a "perfect dataset approach", in which we apply the missingness pattern of the remote sensing datasets onto their matching variables in the ERA5 reanalysis data. Original and imputed values are subsequently compared for benchmarking. Our highly modular approach iteratively produces estimates for the missing values and fits a model to the whole dataset, in an expectation-maximisation alike fashion. This procedure is repeated until the estimates for the missing data points converge. The method harnesses the highly-structured nature of gridded covarying observation datasets within the flexible function learning toolbox of data-driven approaches. The imputation utilises (1) the temporal autocorrelation and spatial neighborhood *within* one variable or dataset and (2) the different missingness patterns *across* different variables or datasets, i.e. the fact that if one variable at a given point in space and time is missing, another covarying variable might be observed and their local covariance could be learned. A method based on simple ridge regression has shown to perform best in terms of results and computational expensiveness and is able to outperform simple "ad-hoc" gapfilling procedures. This model, once thoroughly tested, will be applied to gapfill real satellite data and create an inherently consistent dataset that is based exclusively on observations.