Modelling integrated water vapour with machine learning and meteorological data

Author: Chunyang Gao
Supervisors: Laura Crocetti, Dr. Matthias Schartner, Prof. Benedikt Soja

1 Introduction

Water vapor influences weather, climate, and the water cycle on various scales. Understanding atmospheric water vapor is thus crucial for grasping the complexities of Earth's system. It can be quantified as integrated water vapour (IWV) which is the integrated mass of water vapour in a vertical atmospheric column over a unit area. The project aims to model Integrated Water Vapor with machine learning methods and meteorological data.

2 Data and Methods

- Target: IWV dataset (458 stations)
- Features: meteorological data from ECMWF ERA5 data set (85 features)

Position and time features		Meteorological features	
ϕ	latitude	q	specific humidity (at 37 pressure levels)
sin (λ)	sine of longitude	t	temperature (at 37 pressure levels)
cos (λ)	cosine of longitude	z	geopotential
h	ellipsoidal height	sp	surface pressure
t	reference epoch		
sin (doy)	sine of day of year		
cos (doy)	cosine of day of year		
sin (hod)	sine of hour of day		
cos (hod)	cosine of hour of day		

- XGBoost: tree-based ensemble learning scheme, shallow regression trees as weak learners are combined into a strong learner with gradient boosting
- Hyper parameter tuning based on spatially independent folds and temporally independent folds

Table 1: Performance of different algorithms on the test stations

Methods	XGBoost (station-wise) [kg/m^2]	Lasso (station-wise) [kg/m^2]	XGBoost (time-wise) [kg/m^2]
RMSE	1.86	2.87	1.70
MAE	1.30	2.04	1.05

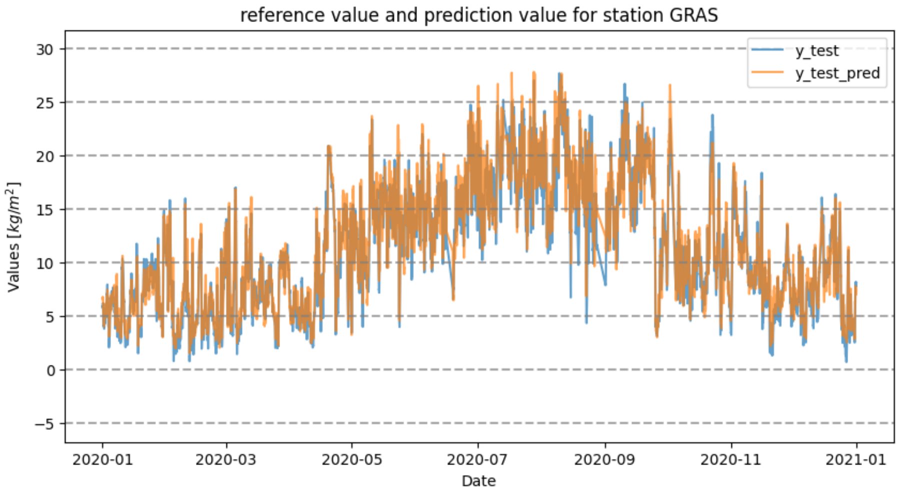
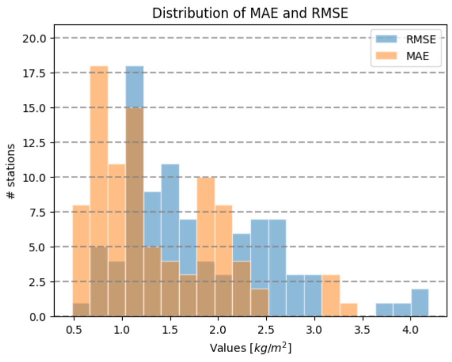


Figure 1: Observations and predictions of one test station with average accuracy

3 Results



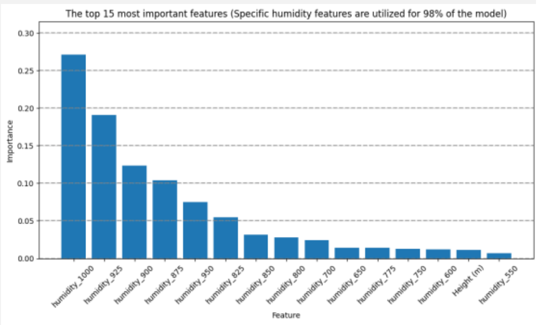


Figure 2: RMSE and MAE for test stations

Figure 3: Feature importance

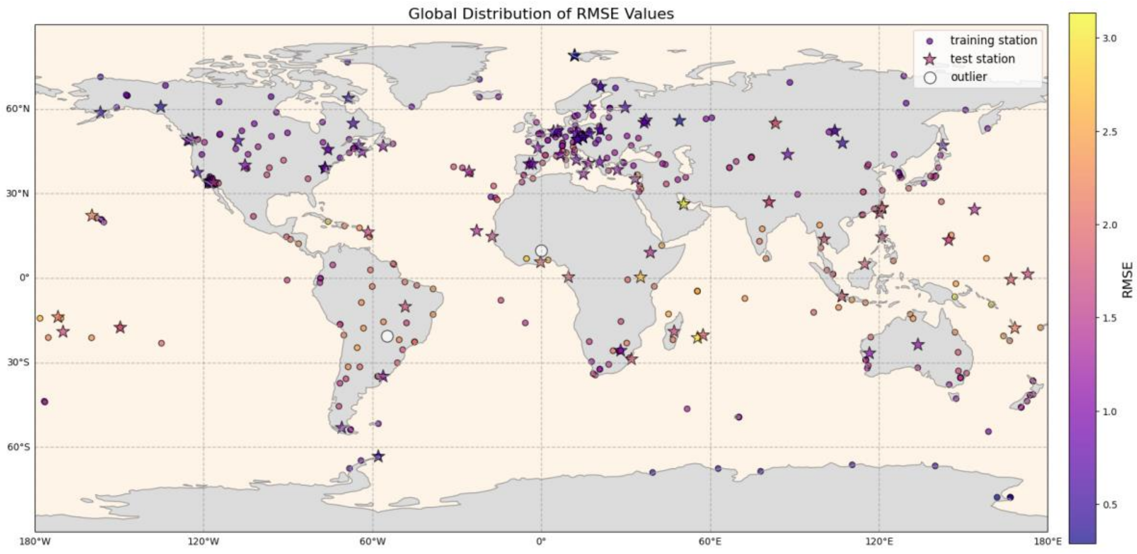


Figure 4: RMSE distribution of training and test stations

4 Conclusion and Outlook

Conclusion

- The method developed by Yuan et al. achieves ± 3.0 kg/m^2 biases with a mean absolute bias (MAB) value of 0.69 kg/m^2. [1] The result of XGBoost is at the same magnitude of precision.
- The model performs well in areas with a dense GNSS station network.
- The humidity at the lower part of the atmosphere is the primary influence factor for IWV, which might explain the model performs well when applied to period outside the training period.

Outlook

- Train the model on multiple years
- Train regional models (different continents)
- Use indirect approach (Modelling IWV by using a global machine learning-based ZWD model)

5 Reference

[1] Yuan, P., Blewitt, G., Kreemer, C., Hammond, W. C., Argus, D., Yin, X., Van Malderen, R., Mayer, M., Jiang, W., Awange, J., and Kutterer, H.: An enhanced integrated water vapour dataset from more than 10 000 global ground-based GPS stations in 2020, Earth Syst. Sci. Data, 15, 723–743, https://doi.org/10.5194/essd-15-723-2023, 2023.