

# Modelling integrated water vapour with machine learning and meteorological data

**Chunyang Gao**

**Supervisors:** Laura Crocetti, Dr. Matthias Schartner,  
Prof. Benedikt Soja

**21.12.2023**



# Introduction

- Water vapor influences weather, climate, and the water cycle on various scales.
- Integrated Water Vapour (IWV): the integrated mass of water vapour in a vertical atmospheric column over a unit area

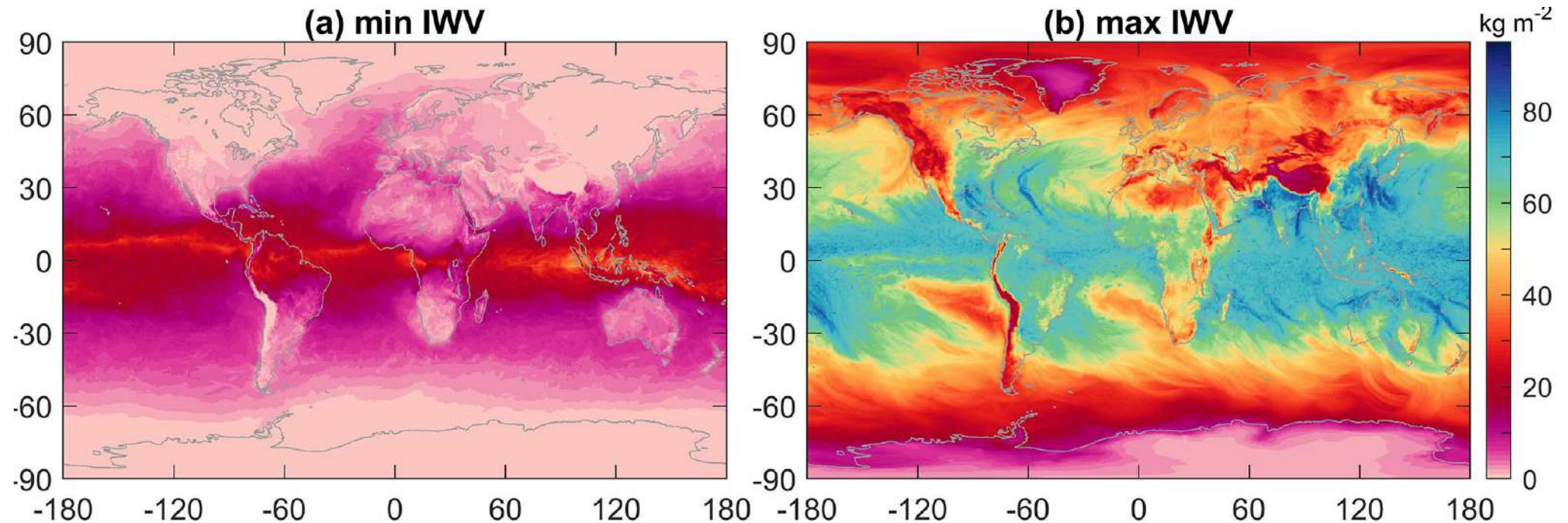


Figure 1: Minimum and Maximum of IWV for 2020 (obtained from ERA5) [1]

# Data and Methods

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

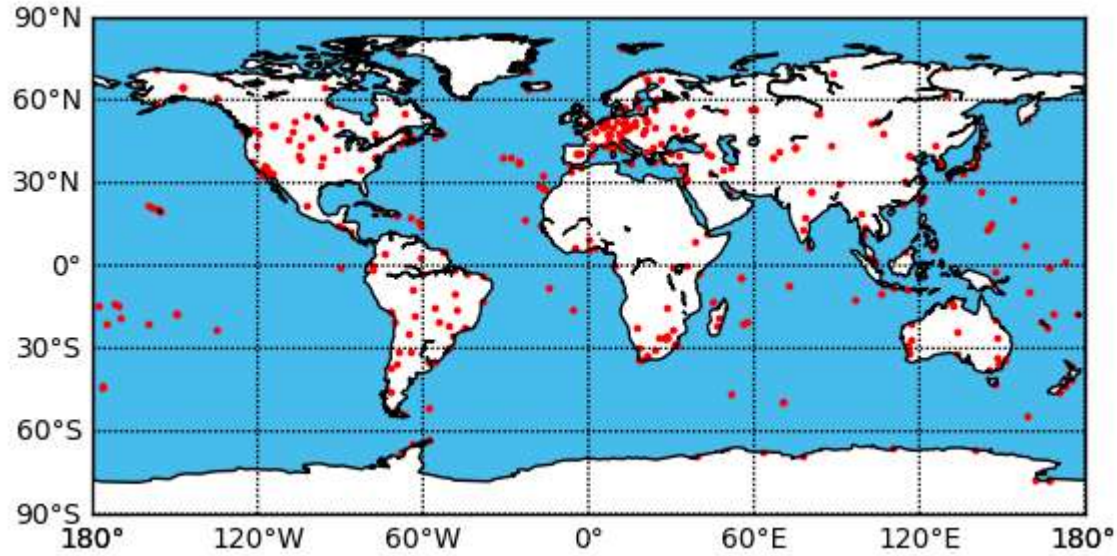


Figure 2: Distribution of GNSS stations

Table 1: Features

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

# Results

Table 2: Performance of different algorithms on the test stations

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

- 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

# Results (for XGBoost station-wise result)

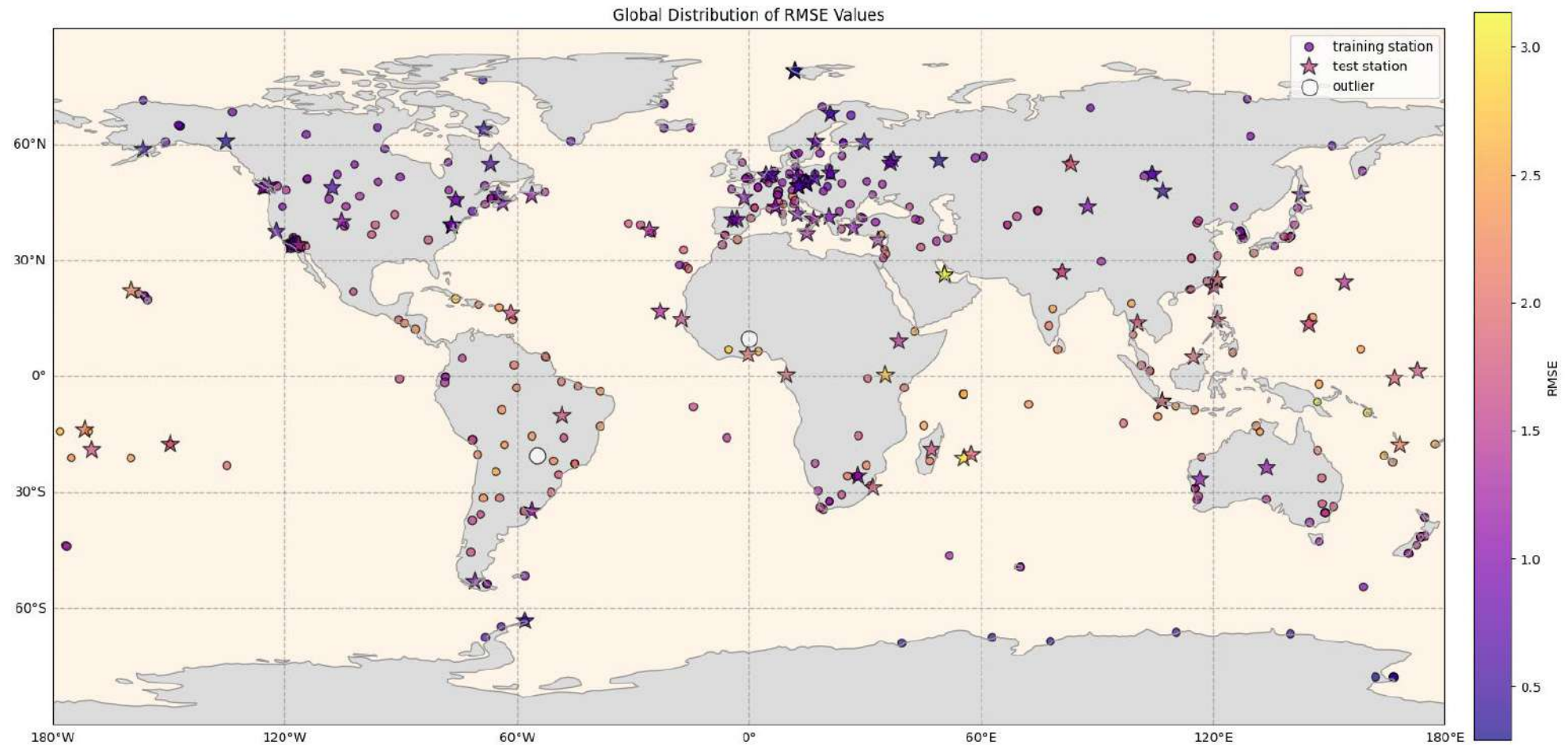


Figure 3: RMSE distribution of training and test stations



# Results (for XGBoost station-wise result)

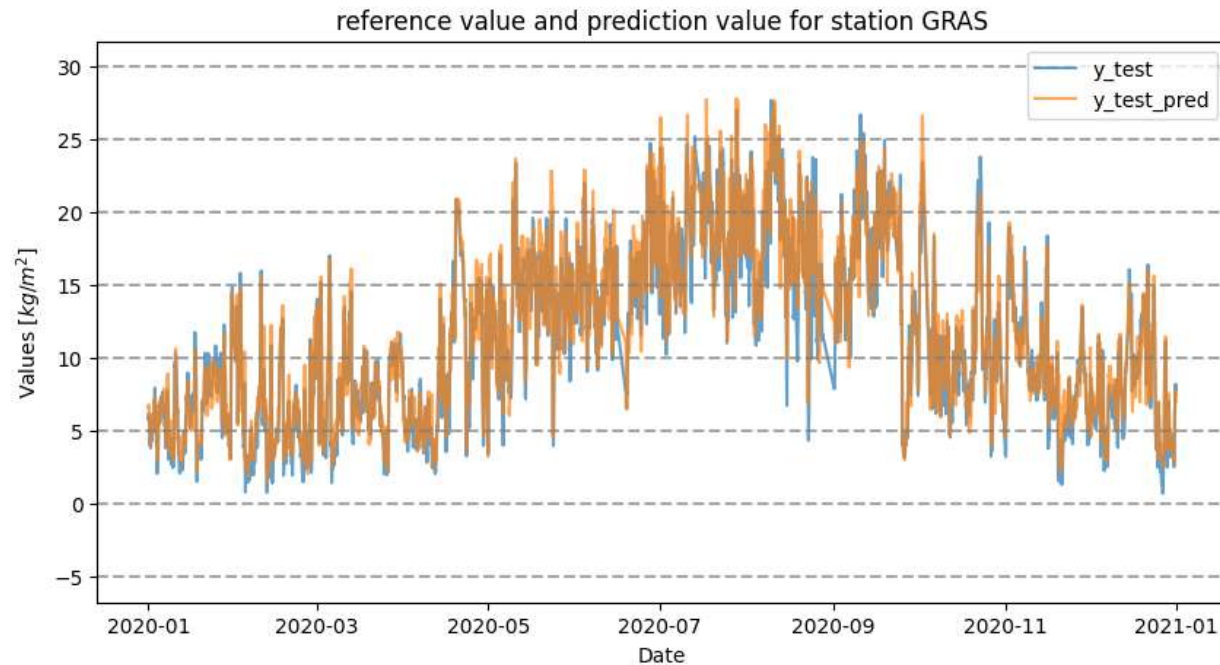


Figure 4: Observations and predictions of one test station with average accuracy

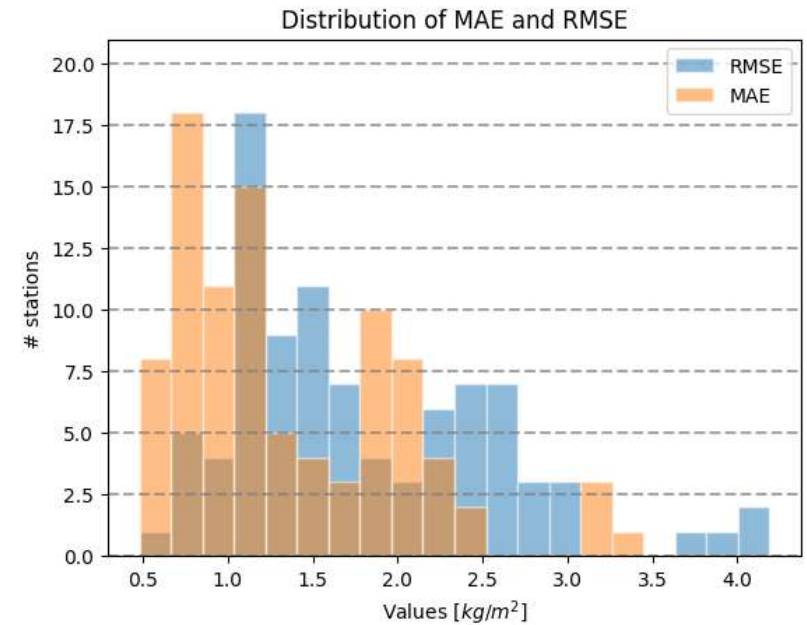


Figure 5: RMSE and MAE for test stations

# Results (for XGBoost station-wise result)

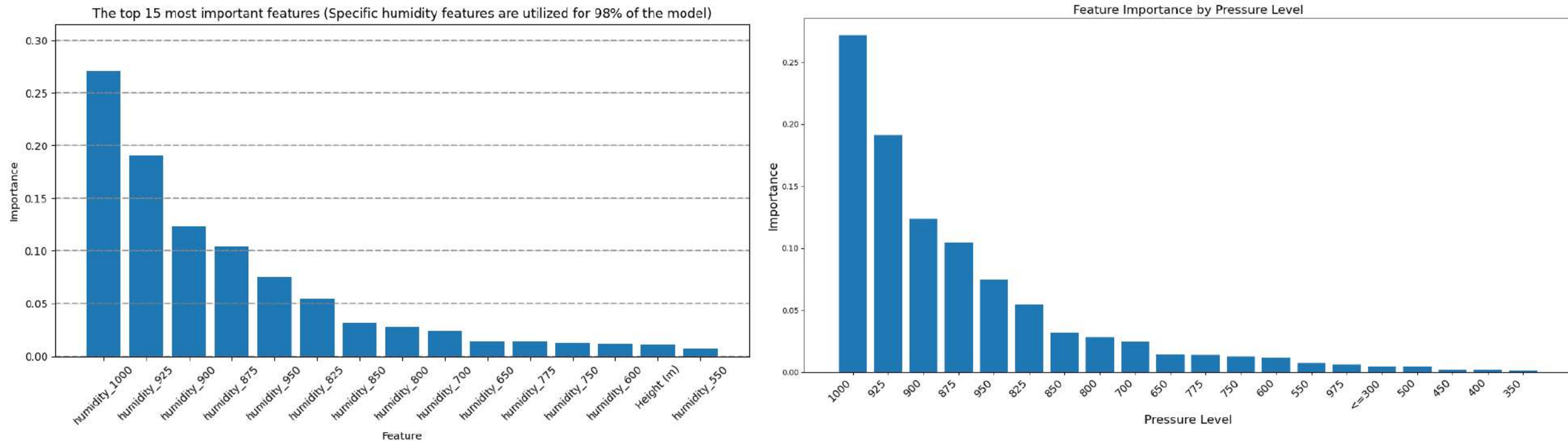


Figure 6: Feature importance

- The humidity at the lower part of the atmosphere is the primary influence factor for IWV

# Conclusion and Outlook

## Conclusion

- The method developed by Yuan et al. achieves  $\pm 3.0$  kg/m<sup>2</sup> biases with a mean absolute bias (MAB) value of 0.69 kg/m<sup>2</sup>. 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)



# 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.



Chunyang Gao  
gaochu@student.ethz.ch

ETH Zürich  
Poster Presentation