Modelling integrated water vapour with machine learning and meteorological data

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SPACE GEODESY

Chunyang Gao

Modelling integrated water vapour with machine learning and meteorological data

Semester Project

Institute of Geodesy and Photogrammetry
Swiss Federal Institute of Technology (ETH) Zurich

Supervision

Laura Crocetti
Dr. Matthias Schartner
Prof. Benedikt Soja

January 2024

Contents

Abstract

Water vapour plays a significant role in earth system. This project developed a model for predicting Integrated Water Vapor (IWV) using observations from 457 GNSS stations. The model takes the geographical location, the time, and a number of meteorological variables as input features. The reference IWV data is the enhanced IWV dataset which is augmented by the high spatiotemporal resolution meteorological information from the ERA5 reanalysis. Unlike traditional instrument-dependent methods, our model leverages solely meteorological variables, positional data, and temporal information to predict IWV, allowing for global applicability beyond the existing GNSS stations. Across all test stations and all observations, the trained model achieves a mean absolute error of 1.30 kg/m^2, respectively a root mean squared error of 1.86 kg/m^2. Comparison of existing method of predicting IWV underlines the good performance of the proposed model. We find the Extreme Gradient Boosting (XGBoost) model outperforms Lasso in predicting IWV. Additionally, we train the model with both spatially and temporally independent data, enhancing its forecasting abilities. The result shows that the model performs better in areas with a denser GNSS network. The error of the time-wise model is lower compared to station-wise model. With regard to feature importance, the specific humidity at the lower part of the atmosphere is the primary influence factor for IWV.

Chapter 1

Introduction

Water vapour is an important component of the atmosphere (Van Baelen et al., 2005). It influences weather, climate, and the water cycle on various scales. Understanding atmospheric water vapour 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 (in units of kg/m^2) (Yuan et al., 2023).

Recently, multiple global Integrated Water Vapour (IWV) datasets have been derived using different methods. Beirle et al. present time series of the global distribution of water vapour columns over more than 2 decades based on measurements from the satellite instruments. Schröder et al. retrieve quantitative information on atmospheric water vapour based on satellite observations and reanalyses. Ferreira et al. derive the distribution and variability of water vapour in the troposphere based on radiosonde measurements from the 1930s to present. Bevis et al. present a new approach to remote sensing of water vapor based on the global positioning system (GPS). These instruments have their own advantages and disadvantages. For instance, radiosondes can provide vertical distribution of water vapour, but their spatial densities and temporal resolutions are low (Yuan et al., 2023). Yuan et al. developed an enhanced GPS IWV dataset by employing accurate meteorological information from the fifth generation of European ReAnalysis (ERA5) with a significantly higher spatiotemporal resolution. Recently, machine learning approaches have also been used to construct models of integrate water vapour. Jadala et al. proposed an IWV prediction algorithm based on the optimized ensemble model. Mei et al. developed a paradigm of deep learning coupling physics and statistics to retrieve water vapour content.

In this work, a ML-based model is trained based on IWV observations of 457 GNSS stations. The reference IWV is taken from the enhanced IWV based on the Nevada Geodetic Laboratory (NGL) and the input features are the geographical location of the GNSS station, as well as the reference time epoch and meteorological variables. The IWV predictions have a good agreement with the enhanced IWV.

Compared to the existing, instrument-dependent IWV retrieval methods, our model has many advantages. First, in contrast to all previously mentioned IWV retrieval methods, our proposed IWV model does not rely on prior IWV or IWV properties to make its predictions. It is based entirely on meteorological variables, position, and time information. This independence allows the model's application globally, beyond the confines of existing GNSS stations. Second, our model uses the enhanced IWV data as the target instead of NGL's operational GPS IWV dataset. Employing meteorological data with a higher spatiotemporal resolution, the enhanced version is generally superior to NGL's operational GPS IWV product because the higher resolution meteorological data can capture small-scale variations of the variables in the complex topographical areas (Yuan

et al., 2023). For the present study, only 457 GNSS stations are utilized to build the model. A much denser GNSS network could be used for IWV predictions with higher-accuracy.

The structure for the report is as follows. In Chapter 2, the reference IWV as well as the meteorological variables are presented. In Chapter 3, the method for the project is presented. The algorithms used, the setup and the validation strategy are introduced. In Chapter 4, the IWV predictions of the final model are shown and compared. Chapter 5 provides the conclusion and outlook of the project.

Chapter 2

Data

2.1 Integrated water vapour

Integrated water vapour (IWV) dataset developed by Yuan et al. is used in the study. It consists of IWV estimates at 12552 GPS stations worldwide in 2020 with a temporal resolution of 5 min. It is derived by using precise GPS tropospheric zenith total delays (ZTD) provided by the Nevada Geodetic Laboratory (NGL), as well as zenith hydrostatic delays (ZHD) and weighted mean temperatures (Tm), obtained from the newly released ERA5 atmospheric reanalysis with a very high spatiotemporal resolution ($0.25^\circ \times 0.25^\circ$, 37 vertical levels, 1-hourly). The mean absolute bias (MAB) of enhanced GPS (enGPS) IWV product at 182 collocated station clusters is only 0.69 kg/m^2 compared with global radiosonde observations. In addition, the enhanced IWV product shows substantial improvements compared to NGL's operational version, and it is thus recommended for high-accuracy applications. It is considered as ground-truth data for the validation of water vapour predictions.

For the present study, we use the IWV of 457 GNSS stations from the year 2020. The distribution of the GNSS stations is illustrated in Figure 2.1. It can be seen that the spatial distribution of these GNSS stations are far from homogeneous. Most stations are located in the northern hemisphere, especially in Europe. The distribution of stations in other continents such as Asia and Africa is very sparse. Only a few stations are available in these areas.

To match the temporal resolution of the meteorological variables, we down-sample the IWV dataset to an hourly resolution by taking the IWV values at every full hour. Finally, we obtain 3561812 data points (8784 hourly time steps $\times$ 457 stations) after removing none values.

2.2 Meteorological variables

The meteorological variables are provided by the European Centre for Medium Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) data set (Hersbach et al., 2020). ERA5 is the fifth generation ECMWF atmospheric reanalysis of the global climate covering the period from 1940 to the present. It is produced by the Copernicus Climate Change Service (C3S) at ECMWF and provides hourly estimates of a large number of atmospheric, land and oceanic climate variables (Crocetti et al., 2023). The data can be accessed through the Climate Data Store2 and are available either as single level data (roughly at surface level) or on 37 pressure levels ranging from 1000 hPa to 1 hPa. For our study, we use specific humidity at 37 pressure levels, temperature at 37 pressure

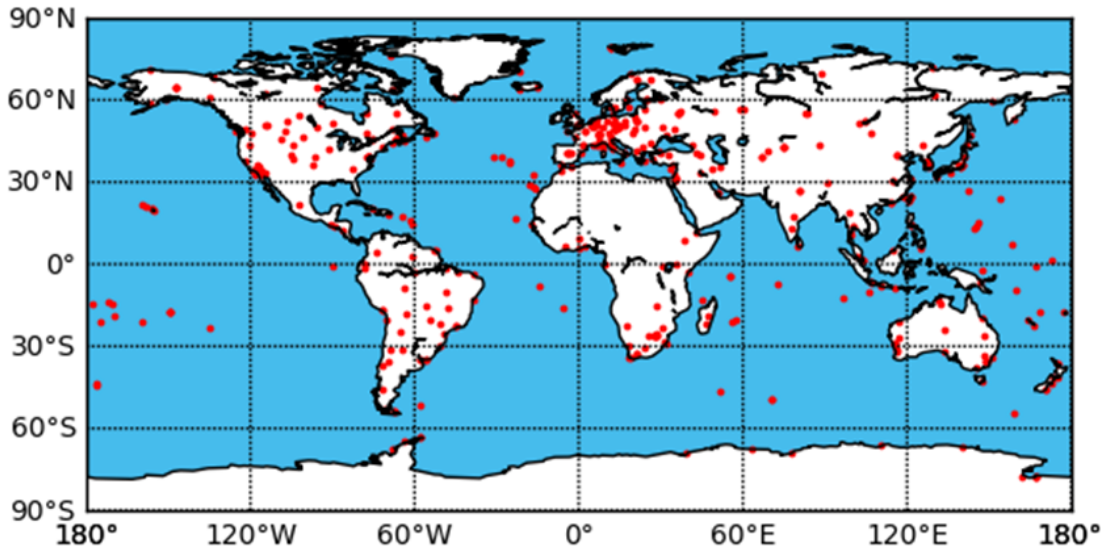

Figure 2.1: Distribution of the 457 utilized GNSS stations

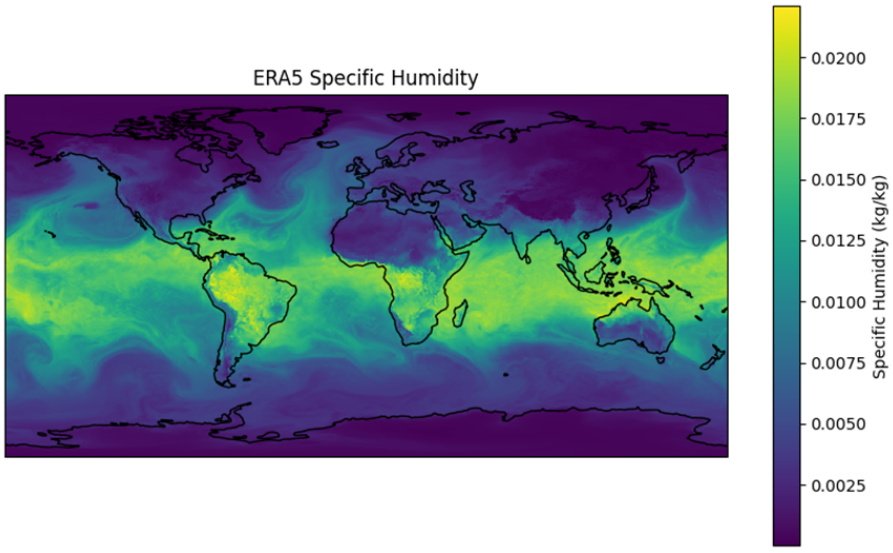

Figure 2.2: Distribution of specific humidity at 1000hPa at 00:00UTC 2020-01-01

levels, geopotential and surface pressure as meteorological variables to train the model. Each grid in the ERA5 dataset has a meteorological value. The meteorological value for a specific station is obtained through interpolation. Then the station information is copied 8784 times (#day of the year $\times$ #hour of the day) to match the timestamps of the meteorological variables in the year 2020. Figure 2.2 shows the distribution of specific humidity at 1000hPa for the first timestamp of the year 2020.

Chapter 3

Methods

To effectively apply machine learning (ML) to determine IWV, three critical issues need to be addressed. Initially, a predictive collection of input features must be selected. Subsequently, an appropriate ML technique must be determined. Finally, the most effective hyper-parameters for the chosen algorithm must be established.

We split the data in two ways to compare the model performance on spatially independent folds and temporally independent folds. First, for spatially independent folds, the 457 GNSS stations are randomly split into 80% training (365) and 20% test stations (92). The test stations are only utilized in the final evaluation. For the experiments to find the best hyper-parameter setup, we rely only on the training stations. At last, training stations are further split into four folds of equal size, four for training and one for model validation. Second, for temporally independent folds, the data points from January to October are used for training and data points of November and December are used for test. In this case, data points for training are split into five folds. First fold includes data points from January to February. Second fold includes data points from March to April. Third fold includes data points from May to June. Fourth fold includes data points from July to August. Fifth fold includes data points from September to October.

In the first investigation, ML algorithms Lasso and XGBoost are tested in the spatially independent folds. The 76 meteorological variables, including specific humidity and temperature at 37 pressure levels, geopotential and surface pressure listed in Table 3.1 are selected, as well as nine position and time variables describing the geographical location of the GNSS station and sample time epoch.

Position and time features		Meteorological features	
ϕ	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		
$\sin(doy)$	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		

Table 3.1: Position, time and meteorological features used in the model

For these two ML algorithms, a hyper-parameter tuning based on grid search is carried out to optimize the predictions on the validation set. Based on the investigation, the XGBoost method is found to be the best model.

Then, the XGBoost is trained on temporally independent folds to find the best hyper-parameter setting for the adapted feature set.

3.1 Algorithms

We test two ML methods performance on the spatially independent folds: Least Absolute Shrinkage and Selection Operator (LASSO) regression and Extreme Gradient Boosting (XGBoost).

Lasso regression is a type of regularized linear regression model. The two main benefits of Lasso regression are feature selection and shrinkage. The Lasso penalty shrinks the less important feature's coefficient to zero. This effectively does feature selection. It also reduces the variance by shrinking the unnecessary variables and makes the selected features more interpretable.

Extreme Gradient Boosting (XGBoost) is a boosting technique that has been designed to optimize distributed gradient boosting (Chen and Guestrin, 2016). It combines weak models to produce a stronger prediction, extreme gradient. It is widely used due to its ability to handle large data sets and achieve state-of-the-art performance in machine learning tasks.

3.2 Setup

As previously mentioned at the beginning of Chapter 3, data from all available GNSS stations are divided into sets for training, validation and test. For each subset, we construct a target vector y and feature matrix.

The feature matrix X, with dimensions of the number of samples by the number of features, is composed of spatial and temporal variables, the geographical coordinates (latitude ϕ, longitude λ), the altitude h of the GNSS station, and the timestamp of the sampling — along with corresponding meteorological variables.

After constructing the feature matrix and target vector, three variables are extracted from the timestamps of each observation, which are in UTC: absolute time as a continuous, real-valued number (t); the day of the year (doy); and the hour of the day (hod). To represent the periodic characteristics of doy, hod and λ, doy and hod are normalized to the range $[0, 2\pi)$ and all three are then transformed to pairs of $\sin(\cdot)$ and $\cos(\cdot)$ values, resulting in two features per variable.

The feature matrix X is normalized before being input into the machine learning algorithms. This is done by deducting the average value of each feature and then scaling the features to have a variance of one.

The data points with missing values of IWV in some timestamps compared to interpolated meteorological values are simply discarded.

3.3 Validation metrics

The two methods used for quantifying the difference between the reference values and predicted values are the root mean squared error (RMSE) and the mean absolute error (MAE). The criterion for hyper-parameter tuning is the RMSE. Data points for test are only used in the evaluation of the final model. For the spatially independent data points, we calculate both RMSE and MAE for

training and test stations. The equations for calculating both of them are as follows.

$$\text{RMSE}_i = \sqrt{\frac{\sum_j (y_{i,j} - \hat{y}_{i,j})^2}{\#samples_i}}$$

$$\text{MAE}_i = \frac{\sum_j |y_{i,j} - \hat{y}_{i,j}|}{\#samples_i}$$

Chapter 4

Results

In this chapter, results for our final, best-performing global model, based on XGBoost trained on spatially independent folds, are presented. We also present the comparisons between different models and algorithms and feature importance.

4.1 XGBoost model trained on spatially independent folds

Parameter	Value
max_depth	7
learning_rate	0.07
n_estimators	130
subsample	0.83

Table 4.1: Hyper-parameters of the station-wise XGBoost model

Table 4.1 shows the hyper-parameters for the final XGBoost model trained on spatially independent folds. The `max_depth` parameter is the maximum tree depth for base learners. The `learning_rate` parameter is the boosting learning rate (shrinks the feature weights after each boosting step to make the boosting process more conservative and prevent over-fitting). The `n_estimators` is the number of gradient boosted trees. The `subsample` parameter controls the fraction of observations used for each tree. A smaller subsample value results in smaller and less complex models, which can help prevent overfitting.

Figure 4.1 shows the spatial distribution of the training stations and test stations' RMSE values. A number of interesting facts can be seen. Test stations with a dense GNSS station network tend to have lower errors. For example, some test stations in Europe and North America have low RMSE values (below 1) while other stations located in South America and Africa have comparatively higher RMSE values. It shows that the predictive of our model is better in areas with a denser GNSS network. There are two outlier stations shown in the graph whose RMSE values exceed 5. These two stations are not located areas with sparse GNSS network. The rationale behind it might be that meteorological parameters might be less accurate. In addition, some meteorological phenomenon might affect the IWV values in the two stations.

Figure 4.2 shows the histogram of RMSE values and MAE values of test stations. We can see that the distribution is skewed towards 0, meaning that most stations have small errors. Most errors

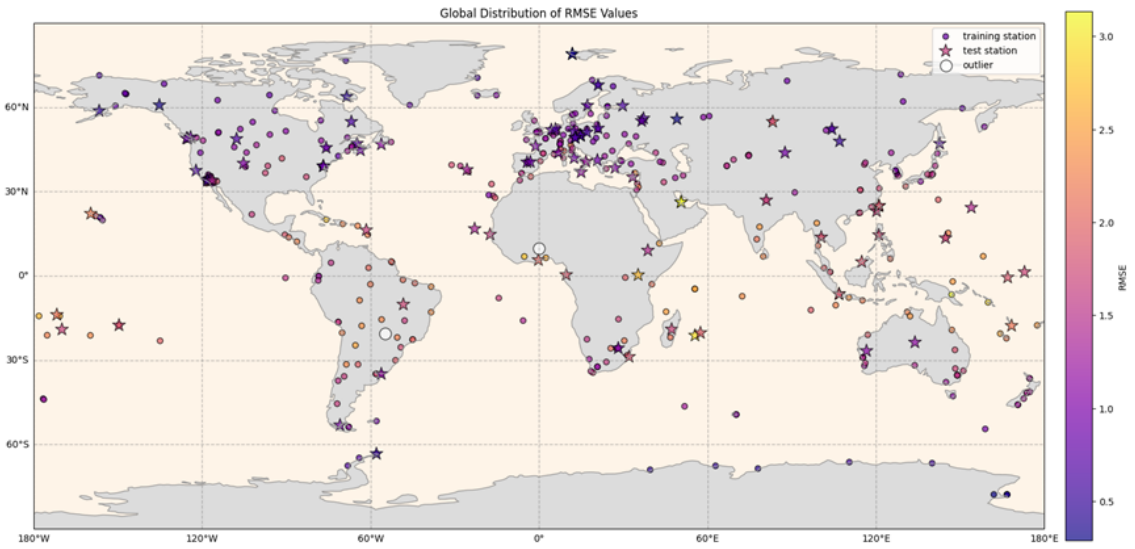

Figure 4.1: RMSE distribution of training and test stations

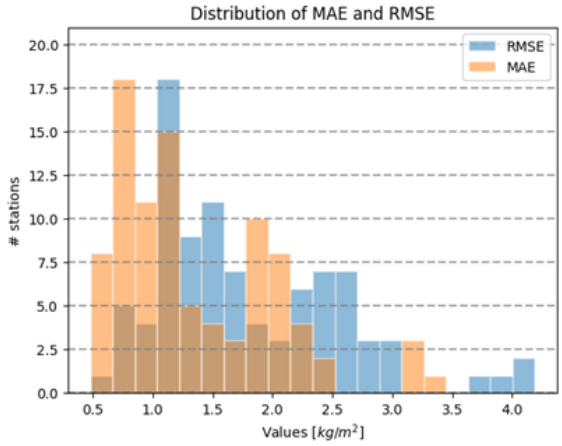

Figure 4.2: RMSE and MAE for test stations

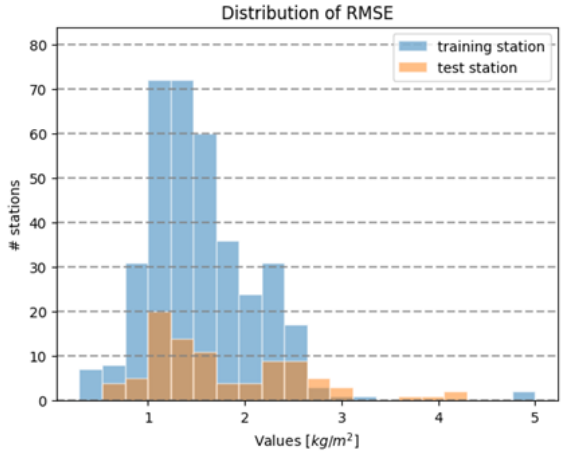

Figure 4.3: RMSE for training and test stations

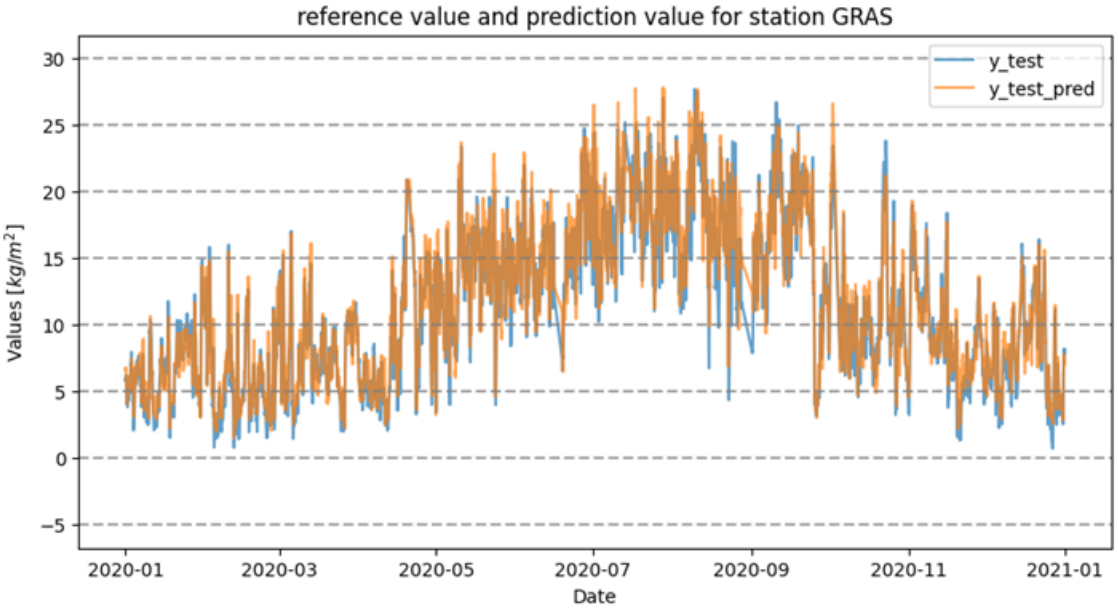

Figure 4.4: Observations and predictions of test station GRAS with average accuracy

are below 4 and concentrated in the range [0.5,1.5], which verifies the predictive skill of our model. Figure 4.3 shows the distribution of RMSE of both training stations and test stations. They have s similar pattern, which means that our model does not over-fit.

Figure 4.4 displays the ground-truth IWV value and predicted IWV value in the year 2020 of an example station GRAS. For every timestamp, the two values are The IWV distribution is very similar for the whole year, which shows that the model developed by us captures the yearly and seasonal variation of IWV well.

4.2 XGBoost model trained on spatially independent folds

Parameter	Value
max_depth	9
learning_rate	0.08
n_estimators	183
subsample	0.80

Table 4.2: Hyper-parameters of the time-wise XGBoost model

Table 4.2 shows the hyper-parameter for the final XGBoost model trained on temporally independent folds.

Figure 4.5 illustrates RMSE of the data points for training and test when splitting the samples into temporally independent folds. From the figure, we can see that the errors of test data points are a little higher than training data points but within a reasonable range. It shows that the model developed by us could perform well when making future predictions. But the clearly higher variability and a sudden jump indicate the temporal over-fitting of the current model to the conditions of the future. It can be explained by high variability of the weather pattern in large parts of the

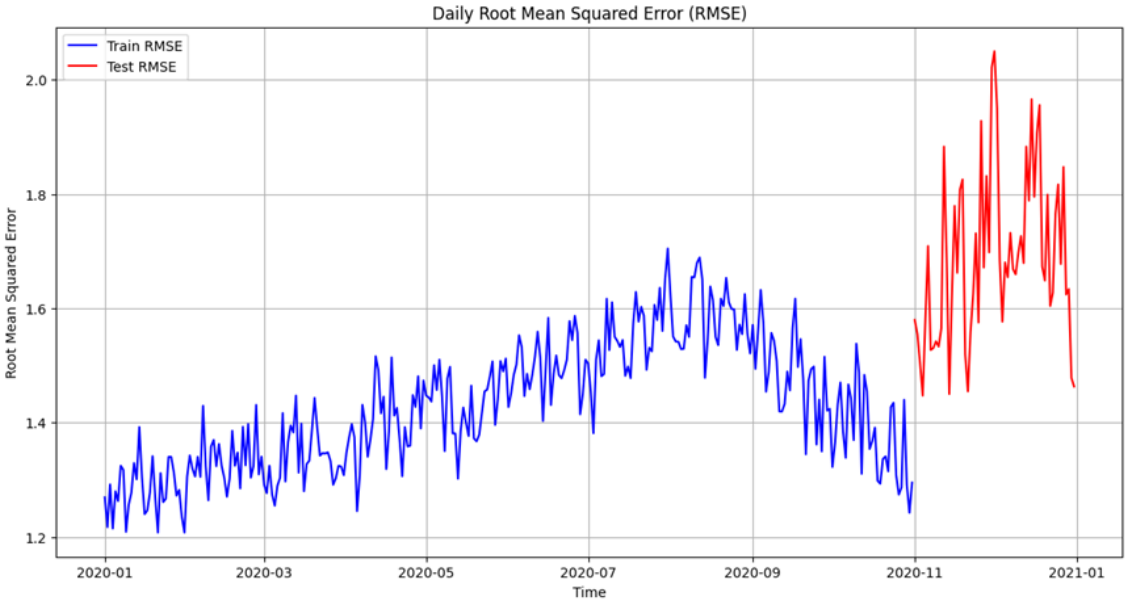

Figure 4.5: Daily RMSE of time-wise XGBoost model

globe in the whole year. We can conclude that data from a specific period are not enough to learn a general model covering the whole research timespan. When training a general model, samples from multiple years are needed.

Compared to the XGBoost model trained on spatially independent folds, it could not capture the seasonal and periodic variability of the IWV.

4.3 Comparisons between models trained on spatially independent folds and temporally independent folds

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

Table 4.3: Comparison of method performance

Table shows the RMSE and MAE of XGBoost and Lasso trained on spatially independent folds and XGBoost trained on temporally independent folds. We can see that XGBoost outperforms Lasso when training the model on spatially independent folds. The IWV dataset 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. The error of our proposed model is at the same magnitude of it, which underlines the good performance of our model.

When using the same XGBoost model, RMSE and MAE values are lower for temporally independent folds (the RMSE of station-wise model is 1.86 kg/m^2 while 1.70 kg/m^2 for the time-wise model). There are several possible reasons for it. First, time-wise XGBoost model has more data points for training than station-wise model, which can lead to a better-performing model. Second,

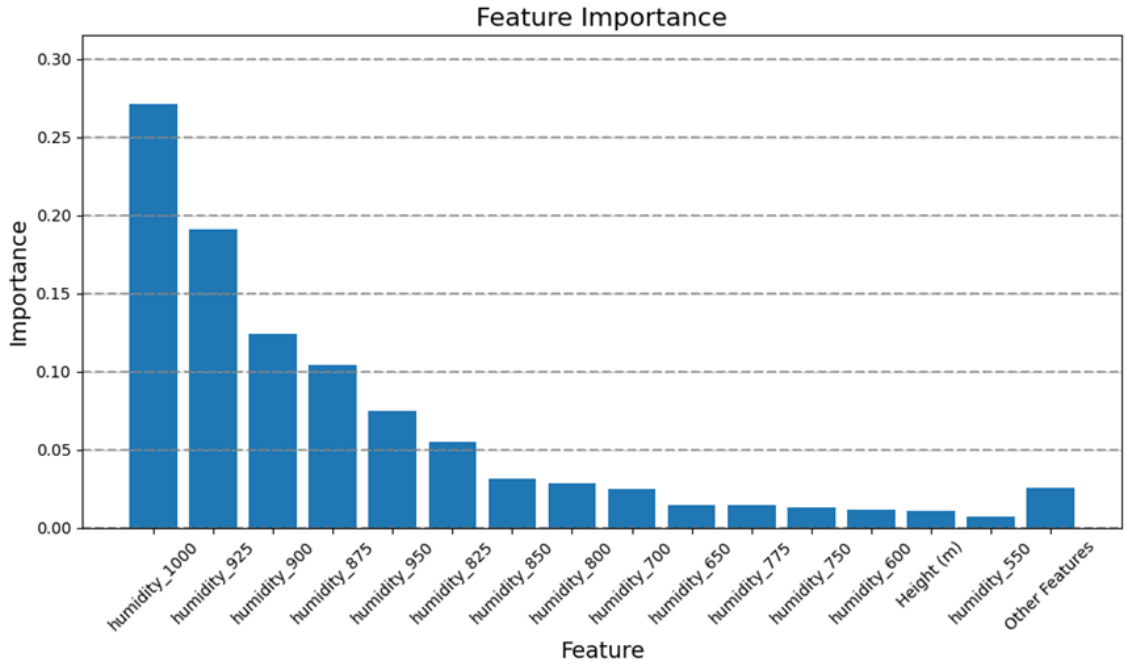

Figure 4.6: The top 15 most important features

we can see from Section 4.1 that there are two outlier stations in the training dataset for the station-wise model. It can limit the predictive ability of the model. While for the time-wise model, it does not have obvious outliers.

4.4 Feature importance

Feature importance represents the relative number of times a particular feature appears in a tree. Figure 4.6 shows the top 15 most important features utilized in the XGBoost model. The other features are grouped in the last bin of the bar chart. Figure illustrates the feature importance in different pressure levels in a descending order

We can see from the figure that the three most important features are the specific humidities at pressure levels 1000 hPa, 925 hPa, and 900 hPa, highlighting that the specific humidity at the lower part of the atmosphere is the primary influence factor for IWV, which also can be seen from Figure 4.9. The specific humidities also correspond to the height of the stations in the dataset, which means the specific humidity in the immediate environment contributes a lot for predicting IWV.

Figure 4.7 and 4.8 show the importance of specific humidity at 37 pressure levels and other features. We can see that specific humidity features are utilized for 98% of the model. In future studies, we could only use the specific humidities to predict the IWV.

Among the position and time features, height plays the most significant role.

Given the fact that the specific humidity accounts for 98% of the model and time features are not so important, it might be the reason why the performance of the time-wise model is not so different from the station-wise model.

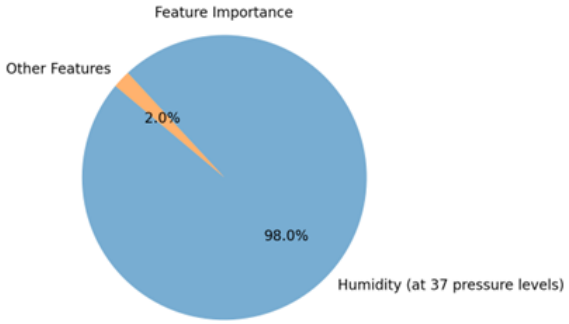

Figure 4.7: The importance raito of the sum of specific humidity and other features

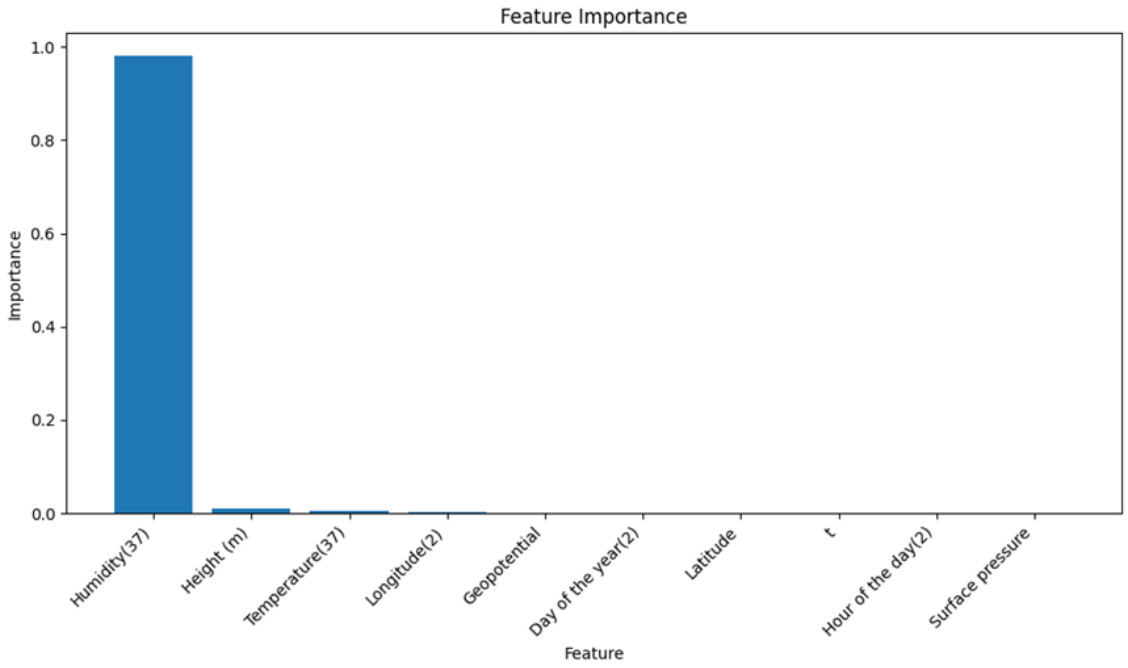

Figure 4.8: Feature importance according to types

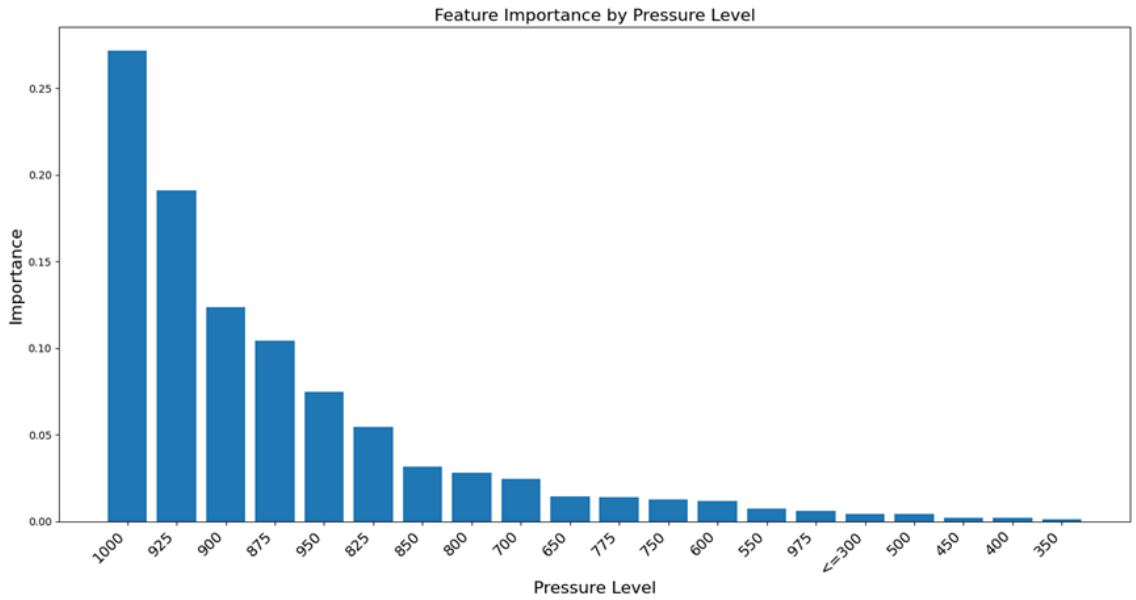

Figure 4.9: Feature importance according to pressure level

Chapter 5

Conclusion and outlook

5.1 Conclusion

In this study, a global ML-based IWV model is presented that achieved a mean absolute error of 1.30 kg/m^2, respectively a root mean squared error of 1.86 kg/m^2. The model utilized the XGBoost algorithm with the geographical location, time epoch, and meteorological variables as its input features. Comparisons between existing method of predicting IWV and ours underline the good performance of the proposed model. The XGBoost is found to be the best-performing model for data points in spatially independent folds. The predictive skill is better in areas with a dense GNSS network. XGBoost performs better in temporally independent folds compared to model trained on spatially independent folds. The reason might be more data points are in temporally independent folds. It is also because that time variables does not play a significant role in the final model. The analysis of feature importance shows that the specific humidity at the lower part of the atmosphere is the primary influence factor for IWV. Specific humidity features are utilized for 98% of the model.

The advantages developed by us are that it is based entirely on meteorological variables, position, and time information, which makes it possible to make predictions globally, beyond the confines of existing GNSS stations. In addition, the use of enhanced IWV also makes its predictions more accurate.

5.2 Outlook

There are several improvement that can be made in the future study.

- In the present study, we only utilize 457 GNSS stations for training the global model. The performance of the model might be limited in areas with less GNSS stations. For further improvement, a much denser GNSS network could be used for higher accuracy.

- The final model developed by us is a global model based on 457 GNSS stations. In addition to creating a global model for the whole year, regional and monthly models could be also generated (Crocetti et al., 2023) to obtain higher accuracy. A trade-off between the global model and specialized models can be determined.

- In a dedicated experiment, the position and time features, the meteorological variables except for the specific humidity could be omitted together.
- Given the ERA5 dataset, the model developed by us can be applied to globally to predict the IWV for every grid at any desired time in the whole global scale.
- Other methods can be used to derive IWV. Signals propagating from GPS satellites to ground-based GPS receivers are delayed by atmospheric water vapor. It is parameterized in terms of a time-varying zenith wet delay (ZWD) which is retrieved by stochastic filtering of the GPS data. The retrieved ZWD can be transformed into an estimate of the integrated water vapor (IWV) (Bevis et al., 1992). Crocetti et al. have developed a global ZWD model based on the Extreme Gradient Boosting (XGBoost). The ZWD predictions made by this model can be transformed into IWV and compared to our model.

Bibliography

Beirle, S., Lampel, J., Wang, Y., Mies, K., Dörner, S., Grossi, M., Loyola, D., Dehn, A., Danielczok, A., Schröder, M., and Wagner, T. (2018). The ESA GOME-Evolution "Climate" water vapor product: a homogenized time series of H_2O columns from GOME, SCIAMACHY, and GOME-2. *Earth System Science Data*, 10:449–468.

Bevis, M., Businger, S., Herring, T. A., Rocken, C., Anthes, R. A., and Ware, R. H. (1992). GPS meteorology: Remote sensing of atmospheric water vapor using the global positioning system. *Journal of Geophysical Research*, 97(D14):15787–15801.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Crocetti, L., Schartner, M., Zus, F., Zhang, W., Moeller, G., Navarro, V., See, L., Schindler, K., and Soja, B. (2023). Global, spatially explicit modelling of zenith wet delay with XGBoost. *Journal of Geodesy*. Submitted.

Ferreira, A. P., Nieto, R., and Gimeno, L. (2019). Completeness of radiosonde humidity observations based on the integrated global radiosonde archive. *Earth System Science Data*, 11:603–627.

Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Jadala, N., Sridhar, M., Ratnam, D., and Tummala, S. (2023). Ensemble based deep learning model for prediction of integrated water vapor (IWV) using GPS and meteorological observations. *Journal of Applied Geodesy*.

Mei, R., Mao, K., Shi, J., Nielson, J., Bateni, S. M., Meng, F., and Du, G. (2023). A novel physics-statistical coupled paradigm for retrieving integrated water vapor content based on artificial intelligence. *Remote Sensing*, 15:4250.

Schröder, M., Lockhoff, M., Fell, F., Forsythe, J., Trent, T., Bennartz, R., Borbas, E., Bosilovich, M. G., Castelli, E., Hersbach, H., Kachi, M., Kobayashi, S., Kursinski, E. R., Loyola, D., Mears, C., Preusker, R., Rossow, W. B., and Saha, S. (2018). The GEWEX water vapor assessment archive of water vapour products from satellite observations and reanalyses. *Earth System Science Data*, 10:1093–1117.

Van Baelen, J., Aubagnac, J., and Dabas, A. (2005). Comparison of Near–Real time estimates of integrated water vapor derived with GPS, radiosondes, and microwave radiometer. *Journal of Atmospheric and Oceanic Technology*, 22:201–210.

Yuan, P., Blewitt, G., Kreemer, C., Hammond, W. C., Argus, D., Yin, X., ..., and Kutterer, H. (2023). An enhanced integrated water vapour dataset from more than 10 000 global ground-based gps stations in 2020. *Earth System Science Data*, 15(2):723–743.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Modelling integrated water vapour with machine learning and meteorological data

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Gao	Chunyang

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zurich, 10, 01, 2024	Chunyang Gao

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.