GNSS-DERIVED PRECIPITABLE WATER VAPOR MODELING USING MACHINE LEARNING METHODS

S. Izanlou*[1], Y. Amerian[1], S. M. Seyed Mousavi[2]

[1]Faculty of Geodesy and Geomatics Engineering, K. N. Toosi University of Technology, Tehran 1996715433, Iran
Saeed.Izanlo@email.kntu.ac.ir; amerian@kntu.ac.ir
[2]School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran
mortezamousavi@ut.ac.ir

Commission IV, WG IV/3

KEYWORDS: GNSS; PWV; Machine Learning; Interpolation; Extreme Gradient Boosting Regression; Random Forest Regression

ABSTRACT:

Atmospheric water vapor plays a vital role in phenomena related to the global hydrological cycle and climate changes, and its Spatio-temporal modeling and prediction help to identify and predict climatic phenomena. Accordingly, in this study, hourly precipitable water vapor (PWV) data sets for 27 stations receiving Global Navigation Satellite Systems (GNSS) observations in one month and machine learning methods were used to estimate PWV. Machine learning methods used in this study 1. Random Forest Regression (RFR) method 2. Extreme Gradient Boosting Regression (XGBR). The root mean square error (RMSE) in PWV estimation with the RFR method (RFR PWV) is 2.42 mm, and in PWV estimation with the XGBR method (XGBR PWV) is 2.75 mm, and the R-squared (R^2) of the RFR method is 0.74, and for the XGBR method, these values are equal to 0.71. The obtained results show the efficiency and accuracy of both models in estimating PWV, which shows that machine learning methods have been able to recognize the behavior and changes of precipitable vapor in a small spatial and temporal interval. Although both ways had high accuracies, the RFR model performed slightly better and had better accuracy than the XGBR model.

1. INTRODUCTION

Water vapor is one of the most essential and abundant greenhouse gases in the earth's atmosphere and keeps the temperature of the earth's surface above the freezing level. Atmospheric water vapor plays an influential role in global weather, climate change and hydrological cycles. Also, this parameter is essential in many atmospheric phenomena such as flood, precipitation, etc. (Bevis et al., 1992; Philipona et al., 2005). This parameter varies significantly in different spatial and temporal scales. Accurate measurement of water vapor and changes in its distribution has become one of the fundamental problems in synoptic, weather forecasting, and climate research. Therefore, the knowledge of the rapid changes in water vapor is essential for analyzing global and regional water vapor distribution (Gendt et al., 2004; Ning et al., 2016; Wong et al., 2015). Meteorologists have provided many parameters to express the water vapor in the atmosphere. Precipitable water vapor (PWV) is one of the most common. If all water vapor in a vertical column of the atmosphere condenses to a cross-section of one cubic meter, the depth of liquid water in this column is called precipitable water vapor.

Since the demand for accurate and real-time weather services has increased, traditional methods such as radiosondes, water vapor radiometers, and solar photometers cannot continuously estimate water vapor with high accuracy and time resolution. Therefore, the demand for having meteorological values with high spatial and temporal accuracy and resolution using the Global Positioning System increased. (Jin and Su, 2020; Kourtidis et al., 2015). Bevis et al. in 1992, first introduced the theory of meteorology with GPS to estimate atmospheric water vapor with the help of GPS-based ground receiver observations (Bevis et al., 1992). This method has been noticed by researchers as a powerful tool in PWV estimation due to its usability in different weather conditions, continuous observations with very high time resolution, low cost, and PWV estimation with an accuracy of about 1-3 mm compared to radiosonde (Bevis, 1994; Foster et al., 2000; Niell et al., 2001; Ning et al., 2016; Van Baelen et al., 2005; Vey et al., 2009; Zhao et al., 2020). After that, Rocken et al. 1993; implemented the Bevis theory method using two GPS receivers located 50 km apart and compared the obtained results with the water vapor radiometer station (WVR); the difference in water vapor obtained from these two methods was about 1 mm. In 1997, Elgered et al. investigated and modeled air mass movement using four years of GPS network observations. The results show a perfect agreement of PWV obtained from GPS with radiosonde and WVR values. From 1997 to 2001, Rocken et al., Emardson et al., and Niell et al. conducted many studies in the field of PWV estimation using GPS, and the satisfactory results of these studies proved the effectiveness of GPS networks in meteorological studies (Niell et al., 2001; Rocken et al., 1997). Gradinarsky et al. in 2002 compared meteorological satellite, radiosonde, and GPS data to observe the seasonal behavior of PWV and found significant trends using seven years of data (Gradinarsky et al., 2002). Grubbs and Jain, in 2017, used nine years of data in Sweden to examine trends in PWV data obtained using radiometers, radiosondes, and GPS. Other researchers also conducted similar studies (Barman et al., 2017; Duan et al., 1996; Gradinarsky et al., 2002; Jin et al., 2009; Vey et al., 2009; Wagner et al., 2006).

The use of machine learning (ML) methods in recent years to estimate an environmental or physical parameter based on its relationship with other factors has made significant progress. ML techniques are good alternatives for analyzing complex biological systems (Kasampalis et al., 2018; Seyed Mousavi and Akhoondzadeh Hanzaei, 2022). The use of the machine learning method is expected to perform well in PWV estimation, but different techniques and different modeling scales can show other performances. In this article, two machine learning methods, Random Forest Regression (RFR) and Extreme

*Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W1-2022-307-2023 | © Author(s) 2023. CC BY 4.0 License.

Gradient Boosting Regression (XGBR), are used to estimate PWV in the American region. Random forest is an ensemble learning method can be used for regression or classification. The XGBR model was developed by Chen and Guestrin and is an advanced and popular algorithm used in ML.

This study is organized as follows: In Section 2, the study area and the data used are presented. How to estimate PWV from GNSS data is also studied in this section. In section 3, RFR and XGBR methods are explained, and also, these methods are applied to GNSS data. Statistical analyzes and comparisons of models are presented in Section 4. Finally, the conclusion is placed in section 5.

2. STUDY AREA AND DATA

2.1 Study Area

The Plate Boundary Observatory (PBO) network stations were launched in 2008 for 3D strain monitoring in North America and Alaska and have since been developed. Some stations in this network have a high rate of observations (1 and 2 seconds). The stations of this network have been used in this study. The studied area is located between the longitudes of -118.6 to -117.6 degrees and the latitudes from 34.4 to 35.4 degrees. Figure 1 shows the distribution of stations used in this article.

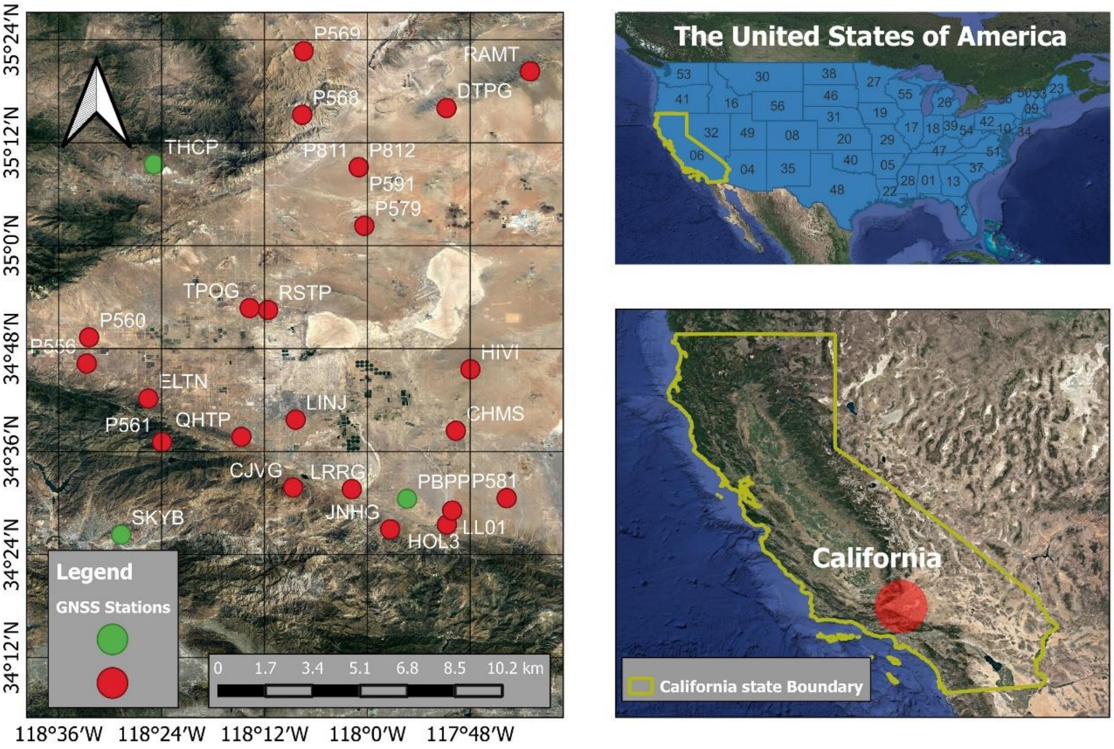


Figure 1. Study area and distribution of GNSS stations

2.2 Data

2.2.1 ECMWF ERA5 data

The fifth generation of Atmospheric Reanalysis (ERA 5) data from the European Center for Medium-Range Weather Forecasts (ECMWF) provides data from 1979 to the present. This database provides parameters such as temperature (T), pressure (P), and other meteorological variables at grid points with horizontal resolution (0.25×0.25 for ERA 5) globally. Due to high spatial and temporal resolution and global coverage, reanalysis products produced by ECMWF have been used in various fields, such as GNSS meteorology. However, the time resolution of the analysis data is different. ERA 5 can provide meteorological data and parameters with a time resolution of one hour. Therefore, ERA 5 has excellent potential in retrieving PWV with high temporal resolution. In this study, surface pressure and temperature data from ERA 5 data series are used.

2.2.2 GNSS PWV data

When GNSS signals pass through the troposphere to reach ground receivers, the signals are delayed. This delay can be converted from oblique mode to zenithal mode using mapping functions; The delay in the zenith direction is called the zenith tropospheric delay (ZTD). ZTD can be divided into two components, tropospheric dry delay (ZHD) and tropospheric wet delay (ZWD). Tropospheric dry delay is a function of pressure and temperature on the earth's surface, which can be calculated using meteorological parameters measured on the earth's surface with an accuracy of a few millimeter (Bevis et al., 1992). In this study, ZHD is calculated using Eq. (1) (Saastamoinen 1973).

$$ZHD = \frac{0.002277 P_s}{(1 - 0.00266\cos(2\varphi) - 0.00000028H)} \tag{1}$$

Where, P_s the surface pressure is in millimeters, φ and H the latitude and orthometric height are in meters, respectively. Due to the extensive time changes of water vapor, it is impossible to model the total delay component with high accuracy.

The delay of the entire troposphere in the zenith direction can be used using accurate GNSS data processing software such as Bernese (Dach et al., 2007). By deducting the tropospheric dry delay from the total tropospheric delay according to Eq (2), we get the tropospheric wet delay in the zenith direction.

$$ZTD = ZHD + ZWD \tag{2}$$

Also, the relationship between ZWD and PWV parameters is defined as Eq. (3).

$$PWV = \Pi(T_m).ZWD \tag{3}$$

In Eq. (3), Π it is the conversion factor, and it is a unitless quantity that is calculated using Eq. (4).

$$\Pi = [10^{-6}(k_3 / T_m + k_2^{'}) R_v \rho_w]^{-1} \tag{4}$$

Where, R_v the gas-specific constant for water vapor is equal to $461.45\, JKg^{-1}K^{-1}$, $k_2^{'}$ and k_3 are the experimental constants and are equal to $17\, Kmbar^{-1}$ and $3.76 \times K^2 mbar^{-1}$ respectively. Also, T_m it is the weighted average of the atmospheric temperature, and it is estimated using the temperature and water vapor pressure of the region (Davis and Herrinch, 1985).

$$T_m = \frac{\int (\frac{e}{T})\, dz}{\int (\frac{e}{T^2})\, dz} \tag{5}$$

Where, e the water vapor pressure is in millibars, and T the temperature is in degrees Kelvin. Experimentally, the value of the conversion factor Π is equal to 0.15 The actual value of this quantity varies between 0.12 and 0.18 depending on the latitude, season and climate of the studied area. In this article, the T_m experimental model presented for the United States, defined as follows, is used.

$$T_m = 70.2 + 0.72 T_0 \tag{6}$$

In Eq. (6), the surface temperature (T_0) is Kelvin.

3. METHODS

As you can see in Figure 2, we first estimated the PWV for the studied period using the GNSS observations and then used the PWV estimated from the GNSS observations and the meteorological parameters that we extracted from the ERA 5 data. We trained machine learning models. In this study, we used 80% of data to train RFR and XGBR methods and 20% of randomly selected data to test and evaluate the obtained model for PWV estimation. By randomly selecting 20% of the data to test the model, we created gaps in the time series, and using the used machine learning models; we estimated the value of PWV in the times when the hole was created. Machine learning methods are explained below.

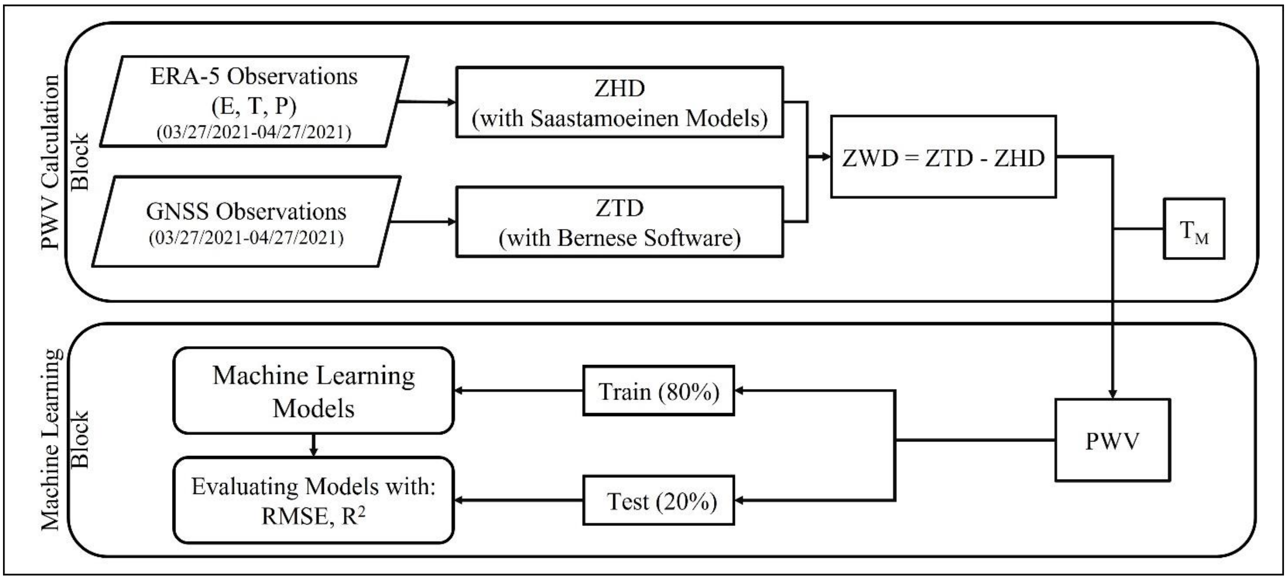


Figure 2. Diagram of different steps of training models.

3.1 Random Forest Regression

RFR is a non-parametric supervised machine learning approach tree-based algorithm where many decision trees are trained with random samples from the training (Shah, Angel et al. 2019).
For regression problems, each tree can consider a large set of regression trees for decision, and Each tree is considered as a vote (Zarei et al., 2021, (Wang, Zhou et al. 2016).

3.2 XGBoost Regression

Extreme Gradient Boosting (XGBoost) is a machine learning regression developed by Tianqi Chen based on a gradient boosting algorithm. it uses residuals to improve the model, mean XGBoost integrates weak regression into strong regression, and iteratively produces new trees to fit the residuals of the previous tree (Jing, Zou et al. 2022).
in addition, in comparison to the prior algorithm first, it can do parallel computing, Second, by using a regularized model it has better management against overfitting (Zamani Joharestani, Cao et al. 2019).
The XGB algorithm can do regression and classification duties in many applications, including remote sensing. The boosting popularity of this algorithm is due to high accuracy and stability relative to other algorithms (Arjasakusuma, Swahyu Kusuma et al. 2020).

4. RESULTS

The evaluation of RFR and XGBR models has been done using the observations of 27 GNSS stations in southwest America. These observations are for days 86 to 117 in 2021. The results of three stations have been randomly selected to analyze the estimation of precipitable water vapor by RFR and XGBR models. Table 1 provides information such as R^2 and RMSE for these stations.

4.1 Comparison of PWV obtained from GNSS and RFR model

After the training stage, it is possible to estimate the amounts of water vapor that can be rained for 86 to 117 days in each station. On the other hand, GNSS PWV values have been estimated for one hour in the studied period in all stations, in the following, the time series of GNSS PWV values of different stations have been compared with the corresponding values obtained from the RFR model. Figure 3 shows the time series of PWV results obtained from GNSS and from the RFR model for selected stations.
The R2 values obtained for the selected stations range from 0.71 to 0.75, and the RMSE ranges from 1.98 to 2.95 mm. As seen in Figure 4, in the time gaps created, the PWV estimate by the RFR model was close to the GNSS PWV values and had a high R2.

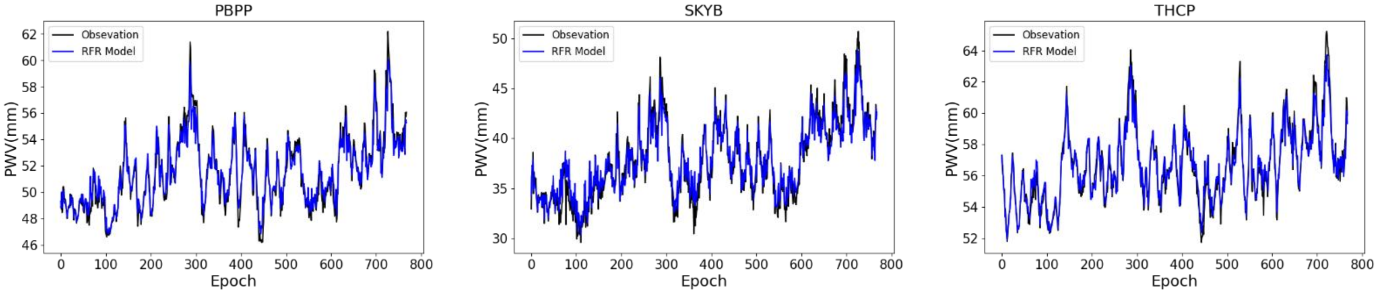
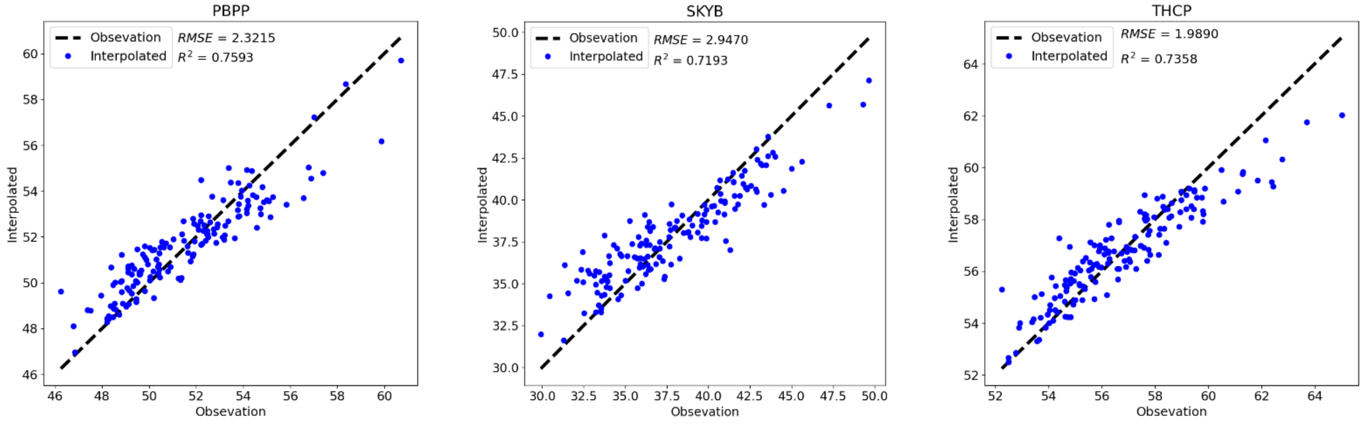


Figure 3. Time series of GNSS PWV and RFR PWV from 27 March 2021 to 27 April 2021 at PBPP, SKYP, and THCP stations



(a) (b) (c)

Figure 4. RMSE and R^2 of the PWV differences for three GNSS stations in the RFR model.

4.2 Comparison of PWV obtained from GNSS and XGBR model

Figure 5 shows the time series of precipitable water vapor estimated using the XGBR model in the desired period, as well as the time series of PWV values obtained from GNSS. As shown in Figure 5, the estimated values for PWV by the XGBR model are very close to the actual values. According to Table 1, the R^2 for the XGBR model includes values ranging from 0.70 to 0.73, and its RMSE has values in the range of 2.23 to 3.49. In Figure 6, the R^2 between the values estimated by the XGBR model and the actual values obtained from GNSS can be seen.

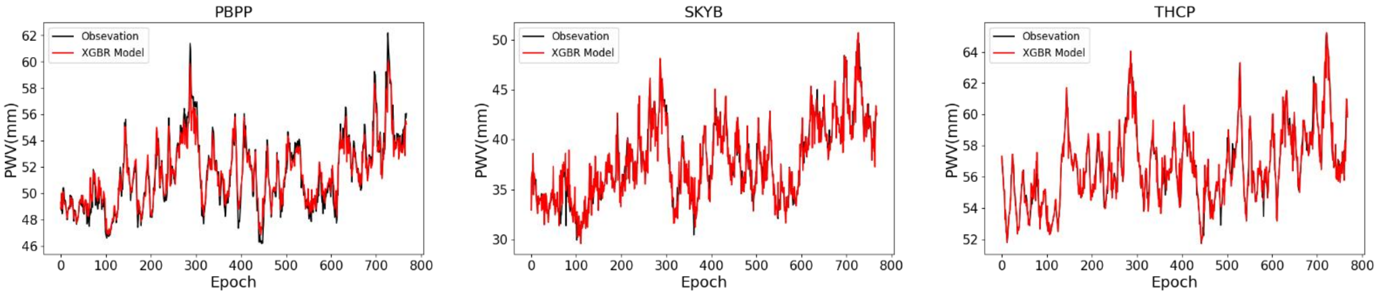
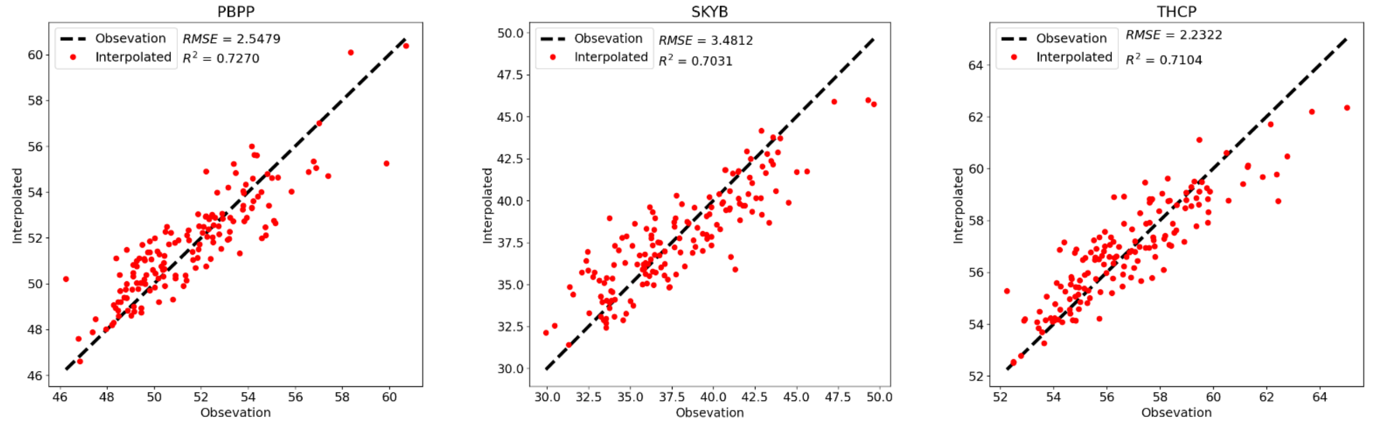


Figure 5. Time series of GNSS PWV and XGBR PWV from 27 March 2021 to 27 April 2021 at PBPP, SKYP, and THCP stations.



(a) (b) (c)

Figure 6. RMSE and R^2 of the PWV differences for three GNSS stations in the XGBR model.

St	Models	Parameters	Train	Test
PBPP	RFR	RMSE (mm)	1.4357	2.3215
		R^2	0.8349	0.7593
	XGBR	RMSE (mm)	1.211	2.5479
		R^2	0.8638	0.7270
SYKB	RFR	RMSE (mm)	2.1305	2.9470
		R^2	0.7951	0.7193
	XGBR	RMSE (mm)	1.7632	3.4812
		R^2	0.8305	0.7031
THCP	RFR	RMSE (mm)	1.2274	1.9890
		R^2	0.8117	0.7358
	XGBR	RMSE (mm)	1.1578	2.2322
		R^2	0.8277	0.7104

Table 1. Statistics of PWV estimates for different models and stations.

5. CONCLUSION

In this study, we used machine learning methods to estimate PWV over a 1 month in the United States based on Random Forest Regression (RFR) and Extreme Gradient Boosting Regression (XGBR). According to the results, the RMSE of the XGBR method varies from 2.23 to 3.49 mm, and the RMSE of the RFR method varies from 1.98 to 2.95 mm. Also, the R^2 for the XGBR method is between 0.70 and 0.73, while the R^2 for the RFR method is between 0.71 and 0.75, these results show the accurate estimation and efficiency of the mentioned methods in water vapor estimation. On the other hand. These two methods can be compared with each other, and according to the results obtained for them, it is easy to understand that the RFR method offers higher efficiency and accuracy

in interpolation than the XGBR method. However, this study only focuses on a small area and global data should be tested in the future. Furthermore, this study only examines the performance of machine learning methods and does not include the evaluation of deep learning techniques, which is left for future work.

6. REFERENCES

Alexandrov, M.D., Schmid, B., Turner, D.D., Cairns, B., Oinas, V., Lacis, A.A., Gutman, S.I., Westwater, E.R., Smirnov, A., Eilers, J., 2009. Columnar water vapor retrievals from multifilter rotating shadowband radiometer data. *J. Geophys. Res. Atmos. 114, 1–28.* https://doi.org/10.1029/2008JD010543

Arjasakusuma, S., et al. (2020). "Evaluating Variable Selection and Machine Learning Algorithms for Estimating Forest Heights by Combining Lidar and Hyperspectral Data." ISPRS International Journal of Geo-Information 9(9): 507.

Barman, P., Jade, S., Kumar, A., Jamir, W., 2017. Inter annual, spatial, seasonal, and diurnal variability of precipitable water vapour over northeast India using GPS time series. Int. *J. Remote Sens. 38, 391–411.* https://doi.org/10.1080/01431161.2016.1266110

Bevis, M., 1994. GPS meteorology: mapping zenith wet delays onto precipitable water. *J. Appl. Meteorol.* https://doi.org/10.1175/1520-0450(1994)033<0379:GMMZWD>2.0.CO;2

Bevis, M., Businger, S., Herring, T.A., Rocken, C., Anthes, R.A., Ware, R.H., 1992. GPS meteorology: remote sensing of atmospheric water vapor using the global positioning system. *J. Geophys. Res. 97, 787–801.* https://doi.org/10.1029/92jd01517

Dach, R., Hugentobler, U., Fridez, P., Meindl, M., 2007. Bernese GPS software version 5.0. User manual. Astron. Institute, Univ. Bern 640, 640.

Davis, J.L., Herrinch, T.A., 1985. • a = • atm dS n (s) - fvac dS z a e (n o 20, 1593–1607.

Duan, J., Bevis, M., Fang, P., Bock, Y., Chiswell, S., Businger, S., Rocken, C., Solheim, F., Van Hove, T., Ware, R., McClusky, S., Herring, T.A., King, R.W., 1996. GPS meteorology: Direct estimation of the absolute value of precipitable water. *J. Appl. Meteorol.* https://doi.org/10.1175/1520-0450(1996)035<0830:GMDEOT>2.0.CO;2

Elgered, G., Johansson, J.M., Rsnn, B.O., Davis, J.L., 1997. hydrostatic [ ] *Ground pressure observed 24, 2663–2666.*

Foster, J., Bevis, M., Schroeder, T., Merrifield, M., Dom, S., Marcus, S., Dickey, J., Bar-sever, Y., 2000. Sea-Level Pressure 27, 2697–2700.

Gendt, G., Dick, G., Reigber, C., Tomassini, M., Liu, Y., Ramatschi, M., 2004. Near real time GPS water vapor monitoring for numerical weather prediction in Germany. *J. Meteorol. Soc. Japan 82, 361–370.* https://doi.org/10.2151/jmsj.2004.361

Gradinarsky, L.P., Johansson, J.M., Bouma, H.R., Scherneck, H.G., Elgered, G., 2002. Climate monitoring using GPS. Phys. Chem. Earth 27, 335–340. https://doi.org/10.1016/S1474-7065(02)00009-8

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma 265, 62–77.* https://doi.org/10.1016/j.geoderma.2015.11.014

Huang, Y., Chen, C., Miao, Y., 2022. Prediction Model of Bone Marrow Infiltration in Patients with Malignant Lymphoma Based on Logistic Regression and XGBoost Algorithm 2022.

Jin, S., Luo, O.F., Gleason, S., 2009. Characterization of diurnal cycles in ZTD from a decade of global GPS observations. *J. Geod. 83, 537–545.* https://doi.org/10.1007/s00190-008-0264-3

Jin, S., Su, K., 2020. PPP models and performances from single- to quad-frequency BDS observations. *Satell. Navig.* 1, 1–13. https://doi.org/10.1186/s43020-020-00014-y

Jing, X., et al. (2022). "Remote Sensing Monitoring of Winter Wheat Stripe Rust Based on mRMR-XGBoost Algorithm." Remote Sensing 14(3): 756.

Kasampalis, D.A., Alexandridis, T.K., Deva, C., Challinor, A., Moshou, D., Zalidis, G., 2018. *Contribution of remote sensing on crop models: A review. J. Imaging 4.* https://doi.org/10.3390/jimaging4040052

Kourtidis, K., Stathopoulos, S., Georgoulias, A.K., Alexandri, G., Rapsomanikis, S., 2015. A study of the impact of synoptic weather conditions and water vapor on aerosol-cloud relationships over major urban clusters of China. *Atmos. Chem. Phys. 15, 10955–10964.* https://doi.org/10.5194/acp-15-10955-2015

Niell, A.E., Coster, A.J., Solheim, F.S., Mendes, V.B., Toor, P.C., Langley, R.B., Upham, C.A., 2001. Comparison of measurements of atmospheric wet delay by radiosonde, water vapor radiometer, GPS, and VLBI. J. *Atmos. Ocean. Technol. 18, 830–850.* https://doi.org/10.1175/1520-0426(2001)018<0830:COMOAW>2.0.CO;2

Ning, T., Wickert, J., Deng, Z., Heise, S., Dick, G., Vey, S., Schöne, T., 2016. Homogenized time series of the atmospheric water vapor content obtained from the GNSS reprocessed data. *J. Clim. 29, 2443–2456.* https://doi.org/10.1175/JCLI-D-15-0158.1

Philipona, R., Dürr, B., Ohmura, A., Ruckstuhl, C., 2005. Anthropogenic greenhouse forcing and strong water vapor feedback increase temperature in Europe. *Geophys. Res. Lett. 32, 1–4.* https://doi.org/10.1029/2005GL023624

Rocken, C., Van Hove, T., Ware, R., 1997. Near real-time GPS sensing of atmospheric water vapor. *Geophys. Res. Lett. 24, 3221–3224.* https://doi.org/10.1029/97GL03312

Saastamoinen, J. (1973). "Contributions to the theory of atmospheric refraction." Bulletin Géodésique (1946-1975) 107(1): 13-34.

Seyed Mousavi, S.M., Akhoondzadeh Hanzaei, M., 2022. Monitoring and Prediction of the changes in water zone of wetlands using an intelligent neural-fuzzy system based on data from Google Eearth Engine system (Case study of Anzali Wetland, 2000-2019). Eng. *J. Geospatial Inf. Technol. 9, 19–42.*

Shah, S. H., et al. (2019). "A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat." Remote Sensing 11(8): 920.

Van Baelen, J., Aubagnac, J.P., Dabas, A., 2005. Comparison of near-real time estimates of integrated water vapor derived with GPS, radiosondes, and microwave radiometer. *J. Atmos. Ocean. Technol. 22, 201–210.* https://doi.org/10.1175/JTECH-1697.1

Vey, S., Dietrich, R., Fritsche, M., Rülke, A., Steigenberger, P., Rothacher, M., 2009. On the homogeneity and interpretation of precipitable water time series derived from global GPS observations. *J. Geophys. Res. Atmos. 114, 1–15.* https://doi.org/10.1029/2008JD010415

Wagner, T., Beirle, S., Grzegorski, M., Platt, U., 2006. Global trends (1996-2003) of total column precipitable water observed by Global Ozone Monitoring Experiment (GOME) on ERS-2 and their relation to near-surface temperature. J. Geophys. Res. Atmos. 111, 1–15. https://doi.org/10.1029/2005JD006523

Wang, L. a., et al. (2016). "Estimation of biomass in wheat using random forest regression algorithm and remote sensing data." The Crop Journal 4(3): 212-219.

Wong, M.S., Jin, X., Liu, Z., Nichol, J., Chan, P.W., 2015. Multi-sensors study of precipitable water vapour over mainland China. Int. J. Climatol. 35, 3146–3159. https://doi.org/10.1002/joc.4199

Zamani Joharestani, M., et al. (2019). "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data." Atmosphere 10(7): 373.

Zarei, A., Hasanlou, M., Mahdianpari, M., 2021. A comparison of machine learning models for soil salinity estimation using multi-spectral earth observation data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 5, 257–263.* https://doi.org/10.5194/isprs-annals-V-3-2021-257-2021

Zhao, Q., Liu, Y., Ma, X., Yao, W., Yao, Y., Li, X., 2020. An Improved Rainfall Forecasting Model Based on GNSS Observations. *IEEE Trans. Geosci. Remote Sens. 58, 4891–4900.* https://doi.org/10.1109/TGRS.2020.2968124