PREDICTING PRECIPITABLE WATER VAPOR USING EXPLAINABLE MACHINE LEARNING TECHNIQUES

Gupta Archit* and Ablimit Aili
Nanyang Technological University, Singapore

1. INTRODUCTION

Accurate prediction of Precipitable Water Vapor (PWV) is essential for numerous applications such as weather forecasting, hydrological modeling, and climate studies. However, conventional methods of determining PWV require expensive instruments, making them impractical in many situations. Existing methods of predicting PWV use data and computation expensive frameworks like Long-Short-Term-Memory (LSTM) networks, artificial neural networks (ANN) and genetic algorithms (GA) [1], [2]. This work presents a novel approach to predicting PWV that demands reduced data input and computational resources, making it more cost-effective and accessible.

The proposed method employs explainable machine learning algorithms to predict PWV with a high accuracy. Specifically, we use tree-based regression models coupled with model interpretation tools to predict PWV from a small set of easily measurable or readily available meteorological variables such as latitude, longitude, elevation, and humidity. The proposed method was trained using meteorological data from a limited number of United States locations and achieved high accuracy on unseen locations. The results suggest that the proposed method can provide a cost-effective and accessible solution for predicting PWV, making it a valuable tool for various applications.

2. DATA

Data on precipitable water vapor along with other spatial information and physical quantities such as relative humidity, solar zenith angle, latitude, longitude, wind direction, and wind speed, among others, was collected from the Physical Solar Model (PSM) version 3 of the National Solar Radiation Database (NSRDB). We collected data from 1998 – 2022 with a spatial resolution of 4km and temporal resolution of 30 minutes from the United States of America (USA). In total, 25 features were extracted from the source.

The 17 features originally extracted from each location of the dataset are *Air Temperature*, *Clearsky DHI*, *Clearsky DNI*, *Clearsky GHI*, *Cloud Type*, *Dew Point*, *DHI*, *DNI*, *Fill Flag*, *GHI*, *Relative Humidity* (RH), *Solar Zenith Angle*, *Surface Albedo*, *Surface Pressure*, *Total Precipitable Water* (PW), *Wind Direction*, and *Wind Speed*. Additional spatial (*Latitude*, *Longitude*, and *Altitude*) and temporal (*Year*, *Month*, *Day*, *Hour*, and *Minute*) features were also collected.

The PSM v3 model included data for 23 years from 3268 unique locations. For each location, there were 402,960 data points, and in total, there were 1,316,873,280 data points. Given the size of this dataset, utilizing it in its entirety was computationally expensive. As such, after empirical observations, we decided to sample 8,395 data points from each location to, on average, capture data from every day for each location chosen. For our training data, we sampled 30 random locations, and for our testing data, we sampled 10 random locations. The decision to choose a small and random set of locations was driven by our motivation of building a robust, inexpensive, and highly generalizable model.

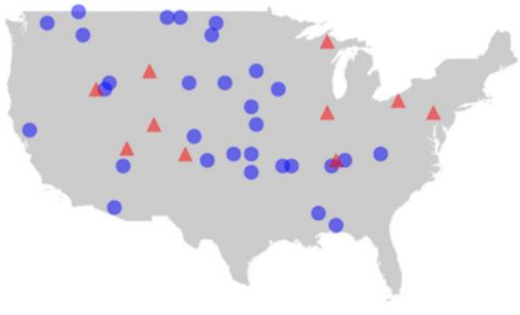

Fig. 1. The train (blue) and test (red) locations

To simplify temporal information, the five temporal features were simplified into three. The *Day (of the Month)* and *Month (of the Year)* features formed the *Day (of the Year)* feature and

**Corresponding author:* Gupta Archit, Nanyang Technological University, Singapore, 637659; e-mail: archit001@ntu.edu.sg the *Minute (of the Hour),* and *Hour (of the Day)* features formed the *Minute (of The Day)* feature.

Moreover, *Specific Humidity* (SH) was also calculated for each data point as a function of Relative Humidity, Temperature, and Pressure [3].

3. METHODOLGY

First, we build a linear regression model to act as the baseline. Then, we employed Random Forest regression and Local Interpretable Model-agnostic Explanations (LIME) [4] sequentially to create a robust, low-cost, and generalizable model.

The choice of Random Forest regression was driven by its capacity to prevent overfitting and yield reliable results without hyperparameter tuning, making it suitable for generalization.

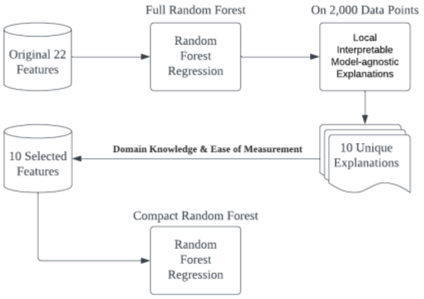

Fig. 2. Model development process

3.1 BASELINE

The baseline model chosen was a single-variate linear regression model with *Specific Humidity* as the independent variable and *Precipitable Water* as the dependent variable.

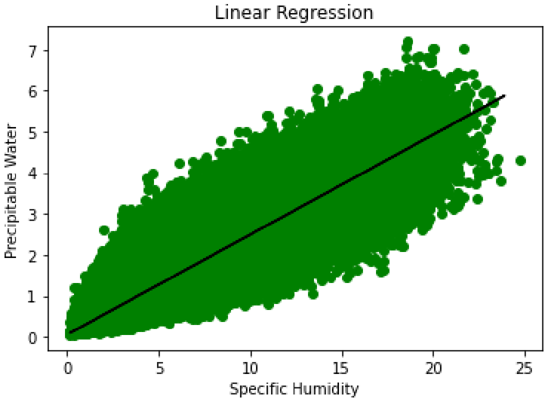

Fig. 3. Regression line plotted on the train dataset

3.2 SP-LIME

To identify relevant features, we built a Random Forest regression model (*FRF*) with all 22 features. As the size of FRF's training sample was 201,480, running SP-LIME on all training instances was impractical. Instead, we empirically chose a sample size of 2,000, and produced 10 unique explanations. These explanations allowed us to calculate the "impact" of each of these variables—either positive or negative—on the final regression.

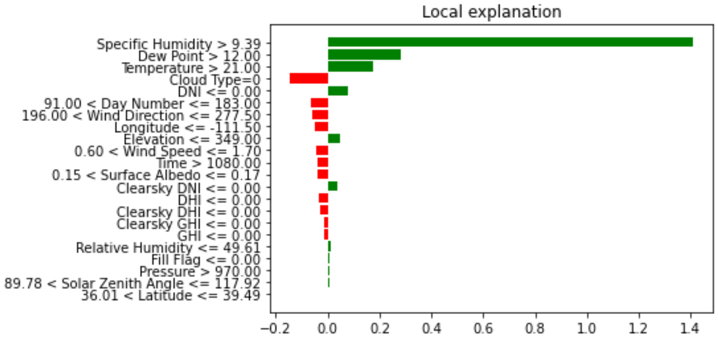

Fig. 4. An example of a positive explanation

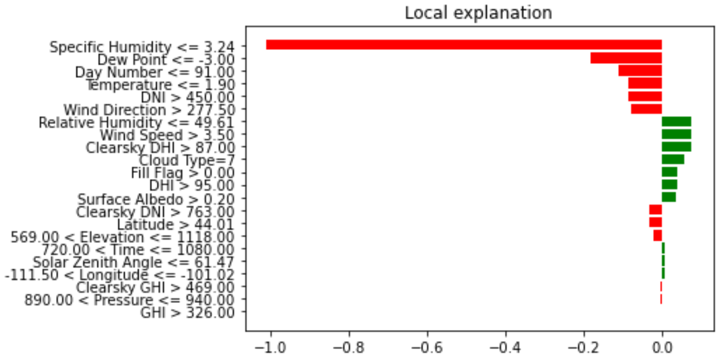

Fig. 5. An example of a negative explanation

These impact values combined with domain knowledge such as ease of measurement and physical relevance led us to choose 10 unique variables that formed the basis of our last and final model, the Compact Random Forest (*CRF*).

3.3 COMPACT RANDOM FOREST

The features chosen for the *CRF* model are *Specific Humidity*, *Dew Point*, *Cloud Type*, *Day Number*, *Fill Flag*, *Wind Direction*, *Latitude*, *Longitude*, *Elevation*, and *Time*. These features were used to train a Random Forest regression model, which not only improved the baseline drastically but also trained much faster than the *FRF* model.

The effectiveness of each of the 10 *CRF* variables was reinforced by Partial Dependence Plots. While *Specific Humidity* had the greatest impact on *Precipitable Water* across the full range of values, the other nine variables greatly influenced the results within their own smaller ranges.

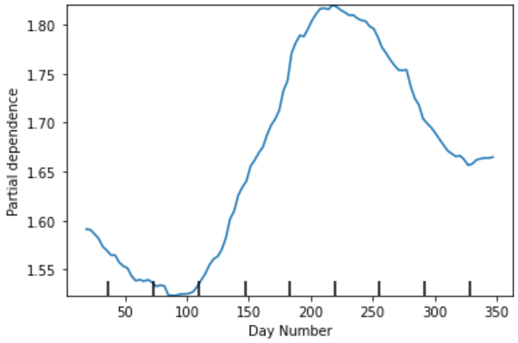

Fig. 6. *Day Number*'s Partial Dependence Plot

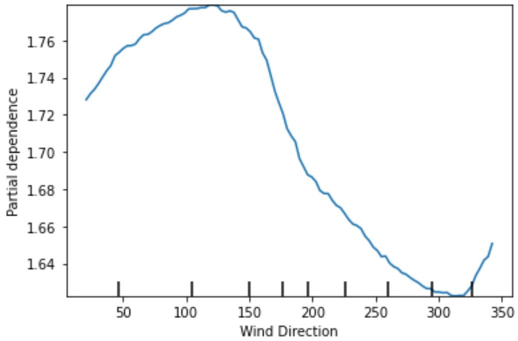

Fig. 7. *Wind Direction*'s Partial Dependence Plot

3.4 OTHER MODELS

For the sake of completeness, other machine learning and physics-based models were implemented and tested. For example, one model was only fed the features used to mathematically compute specific humidity. All these models were outperformed by both *FRF* and *CRF* in either speed, accuracy, or both.

3. RESULTS

To evaluate the performance of each of our models, we use R^2 score as our primary metric. We calculate the mean and standard deviation of the R^2 score across ten random locations to summarize the performance of each model.

While *FRF* outperforms *CRF* in both these metrics, *CRF's* smaller input requirements and shorter run-time makes it our model of choice. Implementing *CRF* on resource-limited cyber physical systems is highly viable given the easily available features it is trained on and the much shorter run time.

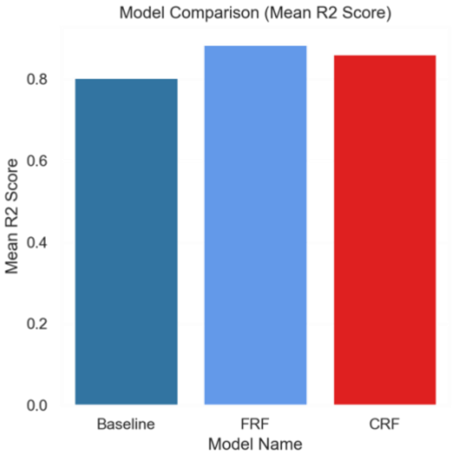

Fig. 8. *Comparison of mean R2 scores of key models*

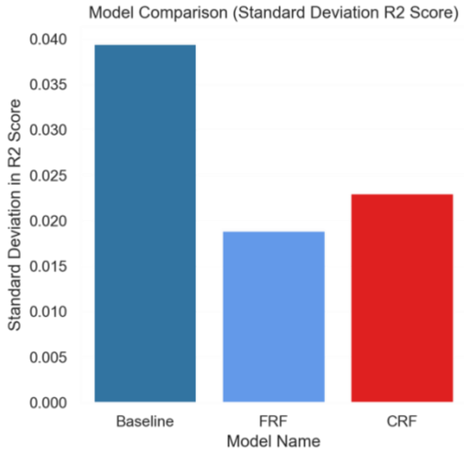

Fig. 9. *Comparison of standard deviation in R2 scores of key models*

5. CONCLUSION

This work presents a novel approach to predicting precipitable water vapor with high accuracy using an explainable machine learning algorithm. The proposed method offers a more cost-effective and accessible solution for PWV prediction, as it requires less data and computational resources than conventional methods. By using tree-based regression models and a small set of easily measurable meteorological variables, the proposed method achieved high accuracy on meteorological data from multiple locations in the United States.

The use of SP-Lime for variable selection contributed to the generalization of our model, which can be applied to meteorological data from other regions of the world. The results demonstrate that the proposed method outperforms the baseline R2 by 7.135%, indicating the effectiveness of the proposed approach.

6. ACKNOWLEDGMENT

We would like to acknowledge the funding and support from Nanyang Technological University for this research project.

7. REFERENCES

[1] M. Jain, S. Manandhar, Y. H. Lee, S. Winkler and S. Dev, "Forecasting Precipitable Water Vapor Using LSTMs," *2020 IEEE USNC-CNC-URSI North American Radio Science Meeting (Joint with AP-S Symposium)*, Montreal, QC, Canada, 2020, pp. 147-148, doi:10.23919/USNC/URSI49741.2020.9321614.

[2] Yue, Y., & Ye, T. (2019). Predicting precipitable water vapor by using ANN from GPS ZTD data at Antarctic Zhongshan Station. *Journal of Atmospheric and Solar-Terrestrial Physics*, *191*, 105059. https://doi.org/10.1016/j.jastp.2019.105059

[3] Aili A, Yin X, Yang R. Global Radiative Sky Cooling Potential Adjusted for Population Density and Cooling Demand. *Atmosphere*. 2021; 12(11):1379. https://doi.org/10.3390/atmos12111379

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org.remotexs.ntu.edu.sg/10.1145/2939672.2939778