A Refining Method of Non-linear Regional Tm Model Based on Random Forest

Qingtong WAN[1,2], Lilong LIU[1,2*], Liangke HUANG[1,2,3*], Wei ZHOU[4],
Yunzhen YANG[1,2], Zixin CHEN[1,2]

1. College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541004, China;
2. Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin 541004, China;
3. GNSS Research Center, Wuhan University, Wuhan 430079, China；
4. Department of Navigation Engineering，Naval University of Engineering，Wuhan 430033，China

KEY WORDS: random forest; nonlinear Tm model; applicability

ABSTRACT:
Weighted mean temperature(Tm) is a critical parameters in GNSS technology to retrieve precipitable water vaper(PWV). By obtaining high-precision Tm, it can provide an important reference data source for regional strong convective weather and large-scale climate anomalies. The high-precision Tm of most areas can be obtained by using the BEVIS model and the surface temperature (Ts). The eastern coastal areas of China are affected by the monsoon climate, which makes the applicability of the method in this area to be improved. The research shows that the Tm which calculated by Fourier series analysis (FTm model) has better applicability in the region than the BEVIS model. However, the method has a single modeling factor, and the precision improvement effect in some area is not obvious. By using the observation data of 13 radiosonde stations in the eastern coastal areas of China from 2010 to 2015. Tm which calculated by numerical integration is used as the reference of the true value. Four of the observation data are selected by the method of random forest (RF). The eigenvalues include the pressure、surface temperature、water vapor pressure and specific humidity are used as input factors. The prediction corrections are added to the deviation of FTm model, and a new Tm is applied to the eastern coast of China which called RFF Tm. Taking the observation data from 2010 to 2014 as the training database, the research area is divided into three areas from south to north according to the latitude. The prediction results of different time scales are studied by the clamping criterion, and then the prediction of random forest is discussed. The correction effect is adaptable in the eastern coast areas of China. The results show that: (1) The RFF Tm model refinement method based on random forest has better adaptability in eastern coastal areas of China, and the applicability of first area is more stable with the prediction time scale than the FTm model. (2) On the time scale with a forecast period of one year, MAE and RMS are 4.7 and 4.6 in third area, 3.2 and 3.8 in second area, and 2.6 and 2.5 in first area. (3) The improvement effect of random forests in the eastern coastal areas of China gradually increases with the prediction period becoming shorter. The predicted deviation values of the eastern coast areas of China reach a steady state when the period is one month. The correction deviations is within 1.5K. The correction range of the third area is better than the second area and first area, which makes up for the shortcomings of the FTm model with low precision in the region. It can be used as a new multi-factor prediction and correction Tm model for GNSS remote sensing water vapor in the eastern coastal areas of China.

1 INTRODUCTION

Atmospheric water vapor is mainly distributed at the bottom of the troposphere, accounting for only 0.1% to 0.3% of the composition of the atmosphere, but it is not only the most active part of the atmosphere, but also one of the important factors affecting the vertical stability of the atmosphere [1-2]. Because the water vapor content has a significant positive correlation with the Precipitable Water Vapor (PWV), atmospheric water vapor content has always been an important research content of weather forecasting and meteorology [3]. At present, commonly used methods for obtaining atmospheric rainfall can be classified into radiosonde, satellite detection, ground-based GNSS, et al. The cost of radiosonde is high and the number of observations is limited; satellite detection is affected by the weather and has many limiting factors; and ground-based GNSS has the advantages of high precision, high spatial and temporal resolution, all-weather, low cost, etc. [4], Therefore, it is widely used.

In the process of inverting PWV using GNSS, the weighted mean temperature is one of the important parameters. At present, the international general calculation method of Tm is the BEVIS model proposed by Bevis in 1992 [5]. It uses the radiosonde station between 27°~65° north latitude to establish a linear model of Tm and Ts. But the applicable Area is smaller, and the applicability in China is poor. The Ref [6-7] combined with multi-factor analysis, found that Tm is periodically negatively correlated with latitude, elevation and pressure (Ps), and is positively correlated with ground temperature (Ts) and water vapor pressure (es). The conclusions were established and a multi-factor regression model for the Chinese Area was established. With the development of GNSS meteorology, the accuracy requirements of Areaal Tm have gradually increased. Various Areaal Tm and Ts models have been established in the Ref [8-10]. However, the previous Areaal models have adopted a linear relationship, and the accuracy in some Areas still cannot meet the application requirements. The Ref [11] based on mathematical statistics model, proved the nonlinear relationship between Tm and Ts, which provides a new direction for the study of Tm. Traditional machine learning methods, such as support vector machine [12], BP neural network [13], Kalman filter model [14], etc., due to the large

* Corresponding author.

proportion of training sample distribution, easy to lead to over-fitting phenomenon and insufficient robustness. As a new machine learning model, random forest can process high-dimensional data samples without dimension reduction processing, and has less parameter debugging and strong versatility, which can effectively avoid over-fitting and has good robust , so it has been widely used in economics, medicine, and exploration [15].

The eastern part of China is affected by the monsoon climate and is prone to strong convective weather, resulting in a significant nonlinear change in the atmospheric weighted average temperature. The Fourier series model can well fit the variation characteristics of Tm in this Area. Therefore, the nonlinear F-Tm model in eastern China is first established based on Fourier series, and the deviation of the model is carried out by random forest method. The prediction is corrected, and finally the nonlinear RFF-Tm model based on random forest is obtained, and the space-time adaptability analysis is carried out on the model.

2 AREA SELECTION AND CALCULATION OF TM

2.1 Research Area

Thirteen radiosonde stations in eastern China were selected as research objects. Because Tm has a large correlation coefficient with latitude, this paper divides eastern China into three research Areas according to 25°N and 35°N. Figure 1 shows the geographical distribution of 13 radiosonde stations. The specific research Area information is shown in Table 1.

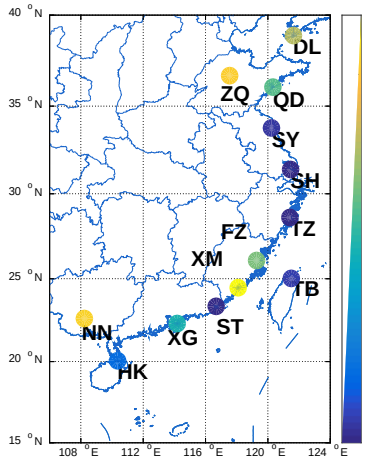

Fig 1. Distribution of 13 radiosonde stations in eastern China

Tab 1.Detailed coordinate table of 13 radiosonde stations in eastern China

Area1	Lon/°	Lat/°	Area2	Lon/°	Lat/°	Area3	Lon/°	Lat/°
Haikou	110.35	20.03	Taibei	121.51	25.03	Qingdao	120.33	36.06
Xianggang	114.16	22.31	Fuzhou	119.28	26.08	Zhangqiu	117.55	36.7
Nanning	108.21	22.63	Taizhou	121.41	28.61	Dalian	121.63	38.9
Shantou	116.66	23.35	Shanghai	121.46	31.4			
Xiamen	118.08	24.48	Sheyang	120.25	33.76			

2.2 Tm calculation method

At present, the commonly used Tm reference value calculation methods mainly include numerical integration method, constant method, BEVIS formula method and approximate integral method. Among them, the numerical integration method has the advantages of high precision and easy implementation, and its calculation result is generally taken as the Tm true value, and the calculation formula is as follows:

$$T_m = \frac{\sum \frac{P_{vi}}{T_i} * \Delta h_i}{\sum \frac{P_{vi}}{T_i^2} * \Delta h_i} \qquad (1)$$

In the formula (1),Ti represents the average atmospheric i-th temperature(the unit is Kelvin), Δhi is the thickness of the i-th layer atmosphere (the unit is m), and Pvi is the atmospheric mean vapor pressure of the i-th layer (the unit is hPa) . Pvi belongs to indirect observation, and is generally calculated by the saturated water vapor pressure calculation formula recommended by the World Meteorological Organization (WMO). The formula is as follows:

$$P_{vi} = 6.112 \exp\left(\frac{17.502 * t}{240.97 + t}\right) \qquad (2)$$

In the formula (2), t is a temperature (the unit is C). In this paper, the measured temperature data of each radiosonde station is extracted, and the Tm obtained by the numerical integration method is taken as the reference value.

3 CONSTRUCTION OF F-TM MODEL AND ITS ACCURACY ANALYSIS

3.1 Fourier series

The Fourier series is a harmonic analysis designed to decompose a function f(x) into the sum of a sine function and a cosine function. For the condition that the function f(x) with a period of 2L satisfies the convergence theorem, the progression of the series can be obtained as:

$$f(\mathrm{x},\mathrm{y}) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(\mathrm{a}_n \cos\frac{n\pi x}{l} + b_n \sin\frac{n\pi x}{l}\right)(\mathrm{x} \in \mathrm{C})$$

$$\begin{cases} a_n = \frac{1}{l}\int_{-l}^{l} f(\mathrm{x})\cos\frac{n\pi x}{l}dx(\mathrm{n}=0,1,2,\ldots) \\ b_n = \frac{1}{l}\int_{-l}^{l} f(\mathrm{x})\sin\frac{n\pi x}{l}dx(\mathrm{n}=1,2,3,\ldots) \\ C = \left\{x \mid f(\mathrm{x}) = \frac{1}{2}\left[f(\mathrm{x}^-) + \mathrm{f}(\mathrm{x}^+)\right]\right\} \end{cases} \qquad (3)$$

3.2 Construction of F-Tm refinement model

In 1992, BEVIS used the BEVIS model (Tm=0.72*Ts+70.2) proposed by the radiosonde station data between 27° N and 65°N in Europe, and is currently used as an empirical model for traditional calculation of Tm.

This paper selects the data of 13 radiosonde stations in China's coastal Areas from 2010 to 2014, and analyzes the impact of water vapor factors on the correlation of Tm-Ts in coastal Areas. The corresponding correlation coefficient (R2) results of each station are shown in Table 2:

Tab 2. Tm-Ts correlation coefficient of 13 radiosonde stations in eastern China

Radiosonde	Dalian	Zhangqiu	Qingdao	Sheyang	Shanghai	Taizhou	Fuzhou
R^2	0.94	0.86	0.93	0.93	0.93	0.92	0.91
Radiosonde	Taibei	Xiamen	Shantou	Nanning	Xianggang	Haikou	
R^2	0.90	0.87	0.84	0.85	0.85	0.7	

It can be seen from Table 2 that the correlation coefficient of Tm-Ts of 13 radiosonde stations in eastern China is between 0.7 and 0.93. The linear equation is not applicable to the calculation of Tm model in the Area. According to the division of latitude, Area 2, Area 3 The correlation coefficient of both is 0.92, while in the lower latitude, the correlation coefficient is only 0.82, and the Haikou radiosonde station is the lowest, which has deviated from the strong correlation range (>0.8). Considering that as the latitude becomes lower, the troposphere and the ionosphere are relatively more active, and the Tm is more active. Therefore, according to the law of Tm change in this Area, the Fourier series model with good applicability is selected and passed through China. The Tm and Ts data of 13 radiosonde stations uniformly distributed in the eastern Area from 2010 to 2014 constructed a nonlinear F-Tm model suitable for the Area. The formula is as follows:

$$FT_m = \text{acos}\left(k * T_s\right) + bsin\left(k * T_s\right) + c \quad (4)$$

and，a=6.943，b=-13.8，k=0.0571，c=275.7.

3.3 accuracy index

In this paper, the deviation (BIAS), the mean absolute deviation (MAE) and the root mean square error (RMS) are selected as the accuracy evaluation factors. The expressions are as follows:

$$\text{BIAS} = \frac{1}{N}\sum\left(X_m^i - X_r^i\right) \quad (5)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|X_m^i - X_r^i\right| \quad (6)$$

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_m^i - X_r^i)^2} \quad (7)$$

Where N is the number of data samples, X_m^i is the model calculated value of the i-th data, and X_r^i is the reference value of the i-th data.

3.4 F-Tm model accuracy analysis

The Ts data of 13 radiosonde stations in 2010-2014 are processed by the F-Tm model and the BEVIS model to obtain the corresponding Tm value. Check the value calculated with the numerical integral, and the deviation result is shown in Figure 2:

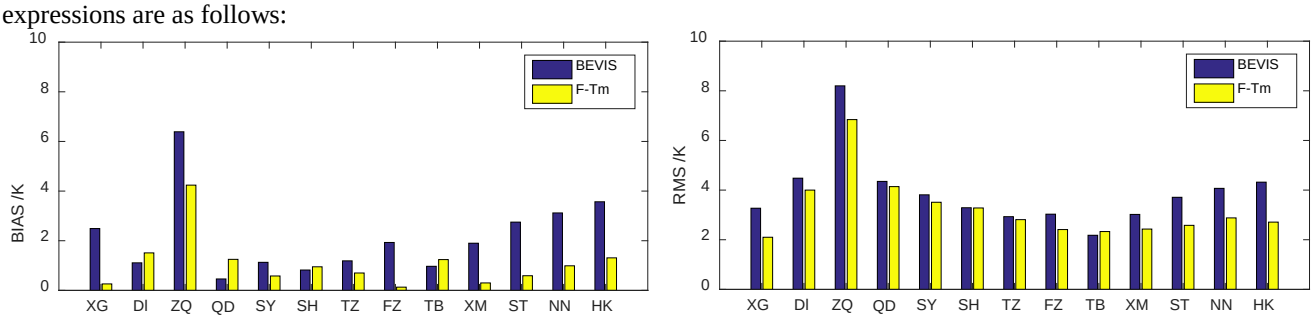

Fig 2. Contrast error of two models of 13 radiosonde stations in 2010-20014

According to the statistics in Figure 2, the F-Tm model has a BIAS value of 0.04k and a RMS increase of 14% in five years. Compared with the BEVIS model, it has better adaptability in eastern China. Among them, the F-Tm model has improved model accuracy in 9 radiosonde stations in Hong Kong, Zhangqiu, Sheyang, Taizhou, Fuzhou, Xiamen, Shantou, Nanning and Haikou; but in Dalian, Qingdao, Shanghai and Taipei. The accuracy of the radiosonde station Area has not improved significantly. Even in a small number of Areas, due to the special geographical location of the Area and industrial pollution, the accuracy of the model has decreased. Therefore, based on the F-Tm model, the deviation is predicted by the random forest method, and the prediction correction number is added to construct the RFF-Tm model:

$$RFFT_m = FTm + \Delta t \quad (8)$$

In the formula, Δt is a prediction correction based on random forest.

4 RFF-TM MODEL CONSTRUCTION AND ITS ACCURACY ANALYSIS

4.1 Random Forest (RF)

Random Forest was proposed by Breiman and Culter in 2001 and belongs to the bagging algorithm in integrated learning. The method divides the data into the original training sample N and the prediction sample Z through the bootstrap resampling technique, and randomly extracts k samples from the N back to generate a new training sample set, and then according to the selection of the feature values, the self-service sample The set generates k classification trees to form a random forest. The two important parameters that have an impact on the model prediction result are the number of decision trees (ntree) and the candidate variable (mtry). The average value of ntree is 1/3 of the number of samples. The classification result of the predicted sample Z is determined by the classification tree. Random forest algorithm has obvious advantages over the over-fitting and complex structure of machine learning methods such as common neural networks and support vector machines. Therefore, it has been widely used in remote sensing image monitoring, ocean subsurface structure prediction, etc.

4.2 Establishment of RFF-Tm model

In this paper, we use the method of random forest, and the deviation value of F-Tm model in 2010-2015 is selected as the data set. The Tm deviation from 2010-2014 is the original training sample. The deviation in 2015 is the training sample, and the selection is related to Tm. The four parameters are used as eigenvalues (pneumatic pressure P, surface temperature Ts, water vapor pressure es, specific humidity s). The ntree value is 3650, the mtry value is 2000, and the correction number in the RFF-Tm model in Equation 8 is obtained. . Compare it with the F-Tm model and the BEVIS model. Select three radiosonde stations for each latitude Area, arranged as follows:

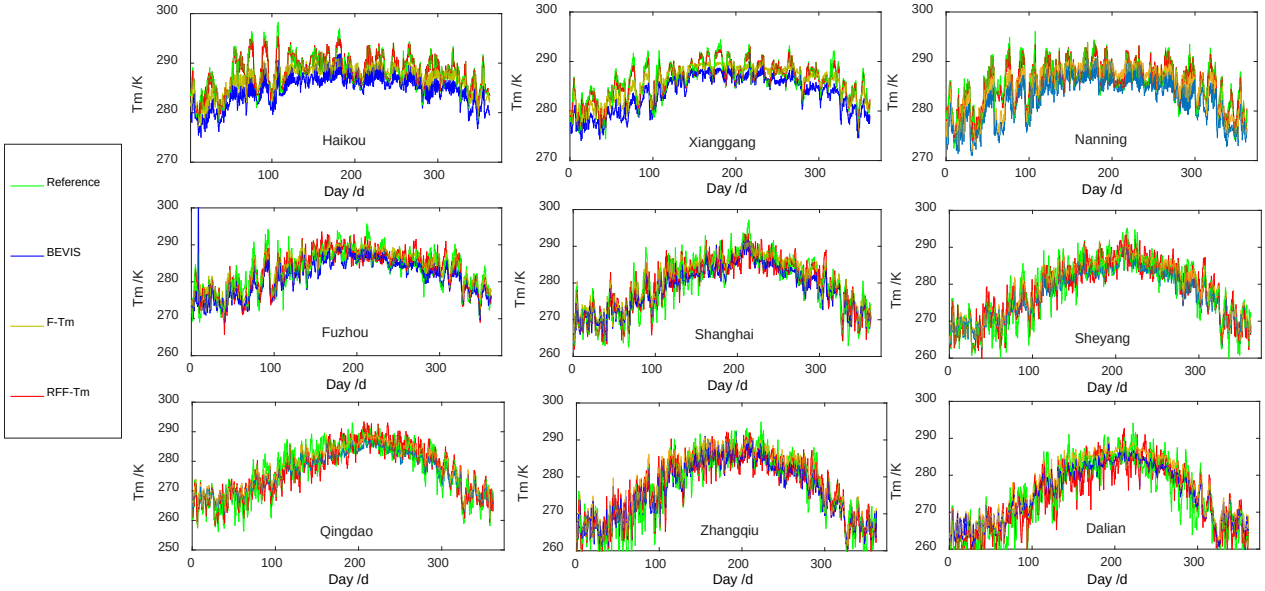

Fig 3. Contrast error of three models in 2015

As can be seen from the figure, the improvement of RFF-Tm in the three radiosonde stations in Area 1 is particularly significant. According to statistics, the MAE of the F-Tm model is increased by 81%, 76%, 77%, respectively. 78%, 72%, and 75% can make up for the tropospheric disorder that appears in the Area due to low latitude. In Area 2 and Area 3, the original accuracy of the F-Tm model is already high, and the bias value used for prediction is small. Long-term machine learning causes distortion of the prediction signal in the Area, and the accuracy is reduced. Therefore, the degree of improvement is not Obviously, the applicability of the RFF-Tm model in these two Areas needs to be further studied.

Based on the premise of not changing the characteristic parameters, ntree and mtry values, the adaptability of the RFF-Tm model in Area 2 and Area 3 is analyzed by adjusting the prediction duration of the random forest. The radiosonde stations of Area 2 and Area 3 are used as random forest prediction models on six time scales of one year, six months, one quarter, two months, one month, and 15 days, respectively. The results of the verification accuracy are shown in the figure below:

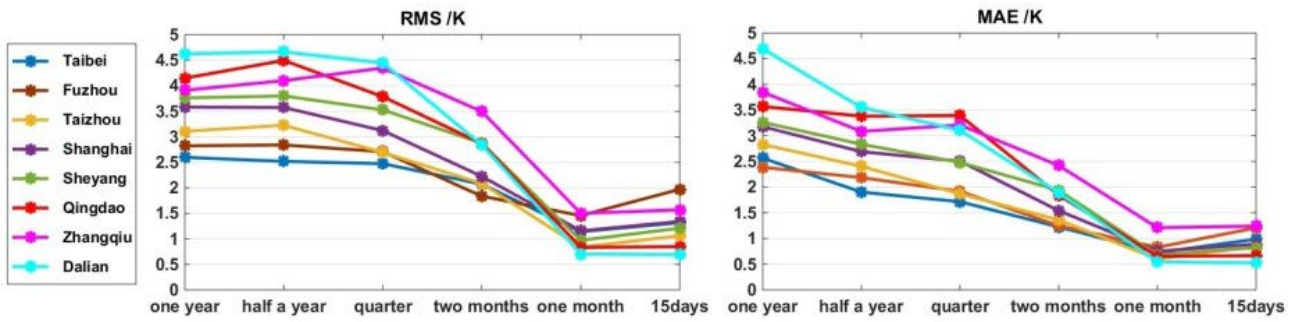

Fig 4. Spatio-temporal adaptability analysis of RFF-Tm model

It can be seen from the figure that the RFF-Tm model has a weaker applicability in Area 3 than Area 2 on the time scale with a forecast period of one year, and MAE and RMS reach a maximum of 4.7 and 4.6 in Area 3, respectively, and the highest in Area 2 respectively. 3.2 and 3.8. As the prediction time scale decreases, the MAE and RMS of the two Areas gradually decrease and tend to be stable, and simultaneously reach a stable value on a one-month time scale, while the prediction accuracy of the time scale of 15 days is not in the Area 3. Significant improvement, in Area 2, a negative growth state. Therefore, in Area 2 and Area 3, the RFF-Tm achieves the best prediction state on the time scale of the predicted time period of one month, and the improvement accuracy of the Area 3 is slightly better than that of the Area 2, and is more stable. Both of them have a good adaptability for MAE and RMS on a one-month time scale. They can be used as a high-precision Tm model for GNSS remote sensing water vapor in eastern China.

5 CONCLUSION

This paper uses the Tm and Ts data of 13 radiosonde stations in eastern China from 2010 to 2014, and uses the Fourier series analysis method to construct the F-Tm model. The results are better than the BEVIS model, but the accuracy in some Areas. There is still room for improvement. Therefore, based on the F-Tm model, four eigenvalues (P, Ts, es, s) are selected by using the random forest method to predict the deviation and obtain the RFF-Tm model. The spatio-temporal adaptability analysis of RFF-Tm is spatially divided into three Areas according to the latitude band, and six time scales are used for random forest prediction. The results show that: (1) The RFF-Tm model has good adaptability in eastern China, and the improvement degree is obvious compared with the F-Tm model. (2) In the Area 1 with low latitude, the RFF-Tm model with a time scale of 1 year has good adaptability and can be applied to long-term sequence analysis. (3) In Area 2 and Area 3, the RFF-Tm model gradually stabilizes with the decrease of time scale, and the time series prediction effect is best in 1 month, and the correction effect of Area 3 is slightly better. In Area 2, and the correction effect is more stable, both are adapted to shorter time series analysis.

ACKNOWLEDGEMENT

This work was sponsored by the National Natural Foundation of China (41664002;41704027);

Guangxi Natural Science Foundation of China (2018GXNSFAA294045;2017GXNSFDA198016;2017GXNSFBA198139);

the “Ba Gui Scholars” program of the provincial government of Guangxi;

and the Guangxi Key Laboratory of Spatial Information and Geomatics (14-045-24-10;16-380-25-01)

REFERENCES

ASKNE J, NORDIUS H. Estimation of tropospheric delay for microwaves from surface weather data [J]. Radio Science, 2016, 22(3): 379-386.

Yao Yibin, GUO Jianjian, ZHANG Bao, et al. A Global Empirical Model of the Conversion Factor Between Zenith Wet Delay and Precipitable Water Vapor[J]. Geomatics and Information Science of Wuhan University, 2016, 41(1): 45-51.

Yu Shengjie, Liu Lintao. Validation and analysis of the water-vapor-weighted mean temperature from Tm-Ts relationship[J]. Geomatics and Information Science of Wuhan University, 2009, 34(6):741-744

YAO Yibin,ZHANG Shun,KONG Jian.Research Progress and Prospect of GNSS Space Environment Science[J].Acta Geodaetica et Cartographica Science,2017,46(10):1408-14020

BEVIS M, BUSINGER S, HERRING T A, et al. GPS meteorology: Remote sensing of atmospheric water vapor using the global positioning system [J]. Journal of Geophysical Research Atmospheres, 1992, 97(D14): 15787-15801.

GONG Shaoqi. The Spatial and Temporal Variations of Weighted Mean AtmosphericTemperature and Its Models in China [J]. JOURNAL OF APPLIED METEOROLOGICAL SCIENCE,2013,24(3):332-341.

WANG Xiaoying, DAI Zaiqiang, CAO Yunchang, et al.Weighted Mean Temperature Tm Statistical Analysis in Ground-based GPS in China. [J]. Geomatics and Information Science of Wuhan University,2011, 36(4): 412-416.

SONG Shuli,ZHU Wenyao, DING Jincai,et al.REAL TIME MONITORING OF PWV FROM SGCAN AND ITS APPLICATIONTEST IN NUMERICAL WEATHER FORECAST[J]. CHINESE JOURNAL OF GEOPHYSICS, 2004, 47(4): 631-638.

ZHANG Hongping,LIU Jingnan, ZHU Wenyao, et al. Remote Sensing of PWV Using Ground-Based GPSData in Wuhan Area[J]. PROGRESS IN ASTRONOMY 2005, 23(2): 169-179.

LV Gepei, YIN Haitao, HUANG Dingfa, at al, Modeling of weighted mean atmospheric temperature and application in GPS/PWV of Chengdu Area.[J]. Science of Surveying and Mapping, 2008, 33(4): 103-105.

YAO Yibin, LIU Jinhong, ZHANG Bao, et al. Nonlinear Relationships Between the Surface Temperature and the Weighted Mean Temperature[J]. Geomatics and Information Science of Wuhan University, 2015, 40(1): 112-116.

LI Meiling HU Yaogai ZHOU Chen,et al. On the short-term Areaal prediction of foF2 based on the support vector machine[J] Journal of Xidian University (Natural Science),2015, 42(5):147-153.

Wang Yong,Zhang Lihui,Yang Jing. STUDY ON PREDICTION OF ZENITH TROPOSPHERIC DELAY BY USE OF BP NEURAL NETWORK[J].Journal of Geodesy and Geodynamics, 2011, 31(3):134-137.

Ghao Y C, Gui X C，Hong Z J，et al. Kalman Filter imaging of ionosphere TFC[J]. Chinese Journal of Geophysics, 2014(11):3617-3624.

Kaminska, JA. A random forest partition model for predicting NO2 concentrations from traffic flow and meteorological conditions [J]. SCIENCE OF THE TOTAL ENVIRONMENT,2018,651(1):475-48