The submitted manuscript describes the establishment of a voxel-based global weighted mean temperature model, named GWMT−D, which is a new version of the GWMT series models. This model obviously improves the modelling performance of weighted mean temperature at higher altitudes compared to the old GWMT models. Such study may be of interested to the community using weighted mean temperature model.

Major comments

01. The primary problem of this manuscript lies in the description and writing. It is found that the methods and explanations in some parts of the manuscript are very difficult to understand. Some descriptions are inaccurate or illogical. The writing of the manuscript needs great improvement and it is recommended to do this with the help of a professional writer.

***Response**: Many thanks. This paper has been revised according to your comments.*

Some examples are shown below.

Page 5, lines 5-6: "Due to the fact that the GGOS data set has been applied in the development GTm−III, the surface Tm from the GGOS data set is also used in the performance assessment of three selected empirical Tm models." Not logical.

***Response**: The GTm−III model was developed using GGOS surface Tm. Thus, this model may have a better consistency with GGOS data than the other models which were not based on the GGOS data. This sentence has been rephrased as '<u>The GGOS data set has been applied in the development of GTm−III and will be also used in the performance assessment of this study</u>' (Page 5 Line 2).*

Page 6, lines 8-10: "This paper takes this feature into account and a new modelling procedure is designed to capture the diurnal variation, i.e. Tm values at any other time are obtained by a spline interpolation method." Not clear.

***Response**: The only purpose of this section is to emphasize the significance of adding the diurnal component in Tm modeling. The details of the modeling will be given in Section 3.2. Therefore, this sentence has been rewritten as '<u>This study takes into account this feature by modelling Tm at each of the four reference times. Whilst, T_m values at any other times can be obtained by the spline interpolation method</u>' (Page 6 Line 5).*

Page 6, lines 26-28: "Long-term Tm time series over the globe can be used for climatological analysis, but its temporal correlation may be too weak to be considered in the Tm modelling process. This suggests that short period of data may lead to an unreliable result." Not logical.

***Response**: Agree. These are two parallel viewpoints but lack of linking words. One the one hand, long-term (>10 yrs) Tm time series may be weakly correlated in the time domain (temporal correlation). On the other hand, a short period (< 1 yr) of data may lead to unreliable results. It has been rephrased as '<u>Long-term T_m data (>10 yrs) over the globe can be used for climatological analysis, but the temporal correlation of T_m time series may be too weak to be considered in T_m modelling. However, a set of short-term T_m data (<1 yr) may be insufficiently for reliable results.</u>' (Page 6 Line 21).*

Page 8, lines 18: "Particularly, the constant-value method performs poorly in both temporal and spatial domains."

***Response**: The constant-value method, which was used in other Tm models for modelling Tm lapse rate, is a possible reason for the poor performance of both GTm−III and GTm_N. The sentence has been moved to Conclusion and rephrased as '<u>The results also confirm that the piecewise linear interpolation of Tm (GWMT−D) is better than the direct modelling of Tm lapse rate (GWMT−IV) and the constant-value method (GTm−III and GTm_N)</u>' (Page 11 Line 7).*

Page 9, lines 4: "terrain of the Antarctic is generally higher than the pressure level of 1000 hPa"

***Response**: Rephrased as '<u>the terrain of Antarctica is generally higher than the height at the 1000 hPa pressure level</u>' (Page 8 Line 23).*

Page 10, lines 9-11: "Comparing with the GTm_N model, better performance of the GTm−III may result from the discrepancy between GGOS surface Tm data (ECWMF reanalysis data) used by GTm-III and NCEP reanalysis data used by GTm_N." Not understand.

***Response**: Rephrased as "<u>Compared with the GTm_N model, the better performance of GTm−III may result from the fact that GGOS T_m, which was derived from ECMWF reanalysis data, is more consistent with the radiosonde data than the NCEP-derived T_m</u>' (Page 9 Line 30).*

02. In the introduction, the GMWT series models are described in details, but no introduction about GMWT-IV model is found here. However, in section 4 this model is compared with the new model GMWT-D.

***Response**: Amended by adding an introduction to GWMT-IV on page 3 line 11: '<u>The GWMT−IV model's T_m lapse rate is a function of geodetic coordinates only</u>'.*

03. Page 2, lines 20-23: Generally, the regression model is also a type of empirical models. Please give a more accurate definition of empirical model.

***Response**: The difference between regression and empirical models is not clear-cut. Generally speaking, empirical Tm model in this study is also a type of regression models. To be strict, 'the regression model' in this paper refers to Bevis formula (Tm-Ts relationship). Therefore, 'regression model' used in this paper has been replaced by 'Bevis formula' in case of confusion (Page 2 Line 21).*

04. Please explain why the four height levels, 0, 2, 5 and 9 km are chosen in the GWMT-D model.

***Response**: The heights are determined empirically. The reason for choosing the 2 km height is that the atmosphere below 2 km suffers the most from the terrain effect. The height level of 9 km, instead of 10 km, is selected because Tm on the 10 km height may be not a valid number (according to the definition of Tm, zero water vapor pressure above this height leads to 0/0). In addition, we found that the result of the four-layer model is satisfactory, and more complex models do not necessarily significantly improve the modeling results. All of these are the main reasons for the selection of the 0, 2, 5 and 9 km height levels in this study (Page 6 Line 14).*

05. Section 3 is hard to understand. Please rewrite this section and describe the procedure of creating GWMT-D model.

***Response**: Amended by polishing. The procedure of developing the new Tm model, which is a multi-dimensional model, is not easy to be well explained. Besides, the separation of relevant figures and tables from the text may also make it difficult to be understood. The general ideas of this section include (1) the types of factors have been considered in GWMT−D and the reasons and (2) the detailed procedure for determining Tm using the new model.*

06. The manuscript claims that the study has improved the diurnal variation in the Tm model. However, it is not found any methods on the diurnal variation modelling in the section "3.1.1 Diurnal variation". In addition, the effect of including diurnal variation in the Tm model is also not shown in the manuscript.

***Response**: Amended. Figure 2 has been redrawn only for the first example at latitude 30 °N, longitude 5 °W, height 2 km since the results of the other two examples are similar. Figures 2(a) and 2(b) were kept the same as before. Figure 2(c) shows the range (max – min) of daily T_m. Figure 2(d) shows power spectrum density of T_m residuals derived from the GTm–III (using daily-mean method) and GWMT–D (using spline interpolation). Figure 2(d) indicates that GWMT–D effectively captures diurnal variations in T_m but GTm−III does not. Although the post-fitting standard deviations of these two methods are very close (~ 3 K) at all the reference times (0, 6, 12, 18 UTC), the spline interpolation used in GWMT–D can significantly remove the diurnal variation (indirect evidence can be seen from Figure 2(d)). However, this can hardly be validated in this study because all the data used are at the reference times. Other data sets (e.g., COSMIC data) may be used in the future studies (Page 6 Line 1).*

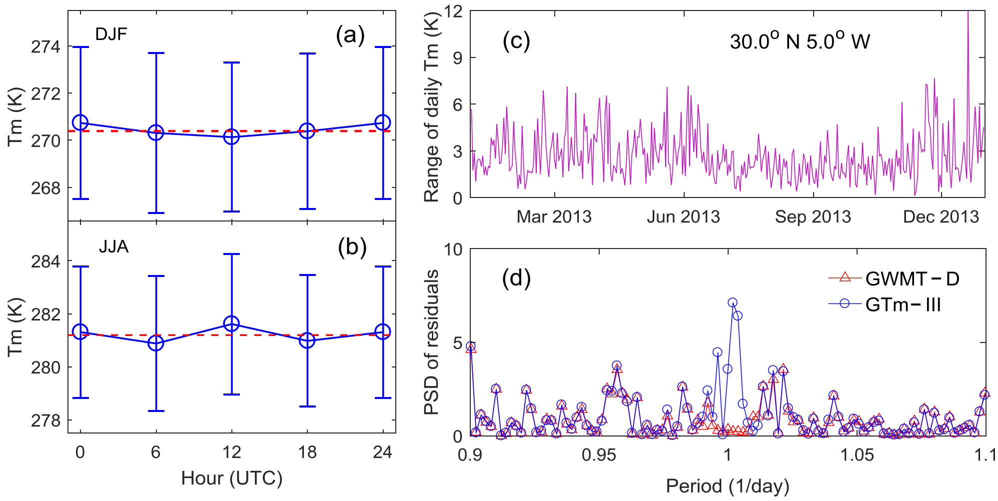

Figure 2. Statistical results of diurnal T_m (mean ± standard deviation) at 2 km height and four reference times during (a) Dec-Jan-Feb (DJF) and (b) Jun-Jul-Aug (JJA) in 2013, (c) the range (max − min) of daily T_m and (d) power spectrum density (PSD) of T_m residuals.

07. Page 8, please explain Table 3.

***Response**: Table 3 shows the means of the biases and RMSs of all global grid points (with a resolution of ~2 degrees), we can summarize the statistical results on each grid. In order to make it easier to be understood, the sentence on Page 8 Line 9 has been rephrased as '<u>As a result, the Biases and RMSs of all global grid points on pressure levels of 925 hPa (~0.6 km) and 600 hPa (~5 km) are given in Table 3</u>'.*

08. Lack of detailed descriptions on the methods used to produce results shown in the Figures 9 and 10 in the manuscript. What is the vertical resolution of Figure 10?

***Response**: Figure 9 is the histogram of the difference between model-derived Tm and radiosonde-derived Tm at all heights from 0 to 9 km. But this figure cannot reflect the terrain's effect. For example, a radiosonde station on the Tibet plateau has a high altitude but it is very close to the surface. Therefore, Figure 10 is for showing the relationship between the accuracy of Tm models and* *the height above surface. These two figures together give sufficient information for the procedure. The vertical resolution of Figure 10 is 100 m.*

09. Please add the number of radiosonde stations in each height interval to Table 5. However, I think there are no radiosonde stations at the altitudes above 5 km.

***Response**: Amended. Only few stations are located higher than 5 km altitude (e.g. the mountainous area and plateau). **However, the % column in this table is the percentage of radiosonde records within the height interval in that of all height intervals.** This description has been added in the title of Table 5 (Page 28).*

Minor comments:

01. Page 3, line 5: "e.g., NCEP-DOE Atmospheric Model Intercomparison 2 (NCEP2) data"?

***Response**: This is the full name of the reanalysis project of NCEP. NCEP2 is the official abbreviation and familiar to researchers, thus this abbreviation keep this abbreviation unchanged but move '(NCEP2)' to the end (Page 3 Line 4).*

02. Page 3, line 27: "2. Data for the determination of Tm" "Data" is enough here.

Response: Amended.

03. Page 4, line 27: "(5) the highest humidity level is far lower than the height of the top troposphere obtained from an empirical model (200~350 hPa)" "far lower" is not accurate. Show the accurate height interval.

***Response**: An empirical model for the height of the tropopause (top layer of the troposphere) used in this study can output the mean height (μ) and standard deviation (σ). If the highest humidity level is less than μ-4σ, this radiosonde data will be removed in the validations. This sentence has been rephrased as 'the highest humidity level (200~350 hPa) is less than $\mu-4\sigma$ where μ and σ are the mean height of the top troposphere and its standard deviation obtained from an empirical model' (Page 4 Line 26).*

04. Page 5, line 2: "(i.e. the lower limit of the integral boundary in Eq. (3) is the surface of the site)". Delete it.

***Response**: Deleted.*

05. Page 7, line 23: "four reference times (i.e. from 0, 6, 12, 18, and 24 UTC) of the day" Delete '24'.

***Response**: Deleted.*

06. Page 9, line 15: "modelling method" may be better.

***Response**: Amended.*

07. Page 13, line 24: "The GTm−II model was identical to GWMT in theory but with different model coefficients" Not clear.

***Response**: Amended. This sentence has been rephrased as "The GTm−II and GWMT models are developed using the same methodology but with different data".*