AAS 23-209

GLOBAL THERMOSPHERIC DENSITY PREDICTION MODEL BASED ON DEEP EVIDENTIAL FRAMEWORK

Yiran Wang[*], Xiaoli Bai[†]

ABSTRACT

The thermospheric density is crucial for calculating the drag of satellites in low-Earth orbit, and the accuracy of the thermospheric density will affect the prediction of satellite trajectories. This paper proposes a global thermospheric density prediction model using a framework based on deep evidence models combining empirical models, geomagnetic and solar indices, and densities inferred from accelerometers from different satellites. In the designed experiments, we study and analyze the model using data from CHAMP, and mixed databases of GRACE-A with different lengths, and test the model on GRACE-A and GOCE. The results show that when the data contains enough data from one satellite and a short period of data from another satellite, the prediction is accurate even when the test case is on a completely new satellite from the training satellite, and the uncertainty estimate is also reliable. The proposed model shows great potential for global thermospheric density predictions.

INTRODUCTION

The thermosphere is a region of the Earth's atmosphere that extends from about 100 km to between 500 and 1,000 km altitude.[1] It is a critical region for space missions, including satellite communication, weather forecasting, and atmospheric studies. However, the thermosphere's density is highly variable and difficult to predict accurately, making it challenging to plan and execute space missions effectively.

Empirical models such as Naval Research Laboratory Mass Spectrometer and Incoherent Scatter Radar Extended (NRLMSISE)[2] and Jacchia-Bowman (JB)[3] models, and physics-based models like Thermospheric General Circulation Models[4] and Global Ionosphere Thermosphere Model[5] have been developed in the past to estimate thermospheric density. While these models have shown some success, they still suffer from uncertain input and boundary conditions in their accuracy, especially during extreme space weather events or in regions with sparse observational data. Additionally, the existing models typically focus on specific geographical regions rather than providing a comprehensive global solution.

To address the limitations of localized models and improve our understanding of the global thermospheric dynamics, there is an increasing demand for a comprehensive global thermospheric den-

[*]Graduate Student, Department of Mechanical and Aerospace Engineering, Rutgers, The State University of New Jersey, NJ, 08854, yw619@rutgers.edu

[†]Associate Professor, Department of Mechanical and Aerospace Engineering, Rutgers, The State University of New Jersey, NJ, 08854, xiaoli.bai@rutgers.edu

sity prediction model. Previous studies have attempted to model the thermosphere density using various methods, including empirical, physics-based, and data-driven models. Emmert et al.[6] analyze a long-term global averaged total mass density database based on two-line element (TLE) data, which covers the year from 1697 to 2007, and the range of altitude is from 200 to 600 km. The author uses many objects (about 5000 objects) and combines them into a single-density database to analyze the density, which reduces the uncertainty in the final computed density and reduces the influence that the objects are with different orbit paths. Weimer et al.[7] computed the total Poynting flux flowing into both polar hemispheres as a function of time and compared its densities in the thermosphere at two altitudes obtained from accelerometers on the CHAMP and GRACE satellites. Elvidge et al.[8] proposed the multi-model ensembles (MME) for forecasting the thermosphere. An MME is a method for combining different, independent models. The main advantage of using the MME is to reduce the effect of model error and bias. Xiong et al.[9] proposed an empirical model called CH-Therm-2018 that uses data from the CHAMP satellite between 2000 and 2009 to predict the thermospheric density. The model utilizes several functions with unknown parameters to obtain the final expression for the density based on latitude, longitude, heights, local time, solar flux index, and season. The model's predictions align well with the CHAMP satellite observations and outperform the NRLMSISE-00 model. Welmer et al.[10] investigated the amplitudes and timings of combined, annual, and semiannual variations in thermospheric neutral density and compared these variations with measurements of the infrared emissions from carbon dioxide and nitric oxide in the thermosphere. The study found a strong correlation between semiannual variations in neutral density and changes in infrared emissions from carbon dioxide and nitric oxide during high solar activity periods.

Several studies have used dynamic reduced-order models (ROM) based on satellite data to improve the thermospheric density prediction accuracy further. Mehta et al.[11] proposed a data-driven methodology to estimate thermosphere composition and temperature simultaneously. The methodology uses modal decomposition to extract a reduced-order representation for the covariance of neutral chemical species and temperature. Gondelac et al.[12] proposed a dynamic ROM for real-time density estimation using TLE data, which outperformed empirical models and had a smaller bias and RMSE. Mehta et al.[13] extended the dynamic ROM to satellite position measurements and demonstrated the effectiveness of the Unscented Kalman Filter in estimating position, velocity, and density. Gondelach et al.[14] further used the ROM with radar and GPS tracking data to estimate density and validated it against accurate SWARM density data. Licata et al.[15] use the ROM to reduce the dimensionality of the Thermosphere Ionosphere Electrodynamics General Circulation Model (TIE-GCM) model and use recurrent neural networks to model the thermosphere with a quicker calculation speed than the numerical model.

In recent years, machine learning methods have recently become increasingly popular in predicting thermospheric density.[16] Bonasera et al.[17] use the Monte Carlo method and deep ensembles to estimate the thermospheric density and the uncertainty from 2002 to 2021. The network is designed to use density data from CHAMP, GRACE, GOCE, SWARM-A, and SWARM-B, the orbital information, and solar and geomagnetic indices as input. Richard et al.[18] build the model based on the Principal Component Analysis (PCA) method and test the density along the satellite orbit from 2002 to 2010.

Our previous studies[19–21] have proposed and enhanced the density estimation framework that integrates information from empirical models JB-2008 and NRLMSISE-00, environment conditions, and satellite measurement data on predicting thermospheric density. In a recent study,[22] we have

successfully demonstrated the efficacy of the deep evidential model in locally predicting thermospheric density based on data from the CHAMP satellite, achieving high accuracy and quality uncertainty estimation. Building on this foundation, our current work aims to extend this approach to build a global thermosphere density prediction model based on deep evidential model. The proposed model considers the data's uncertainty and the model itself, which is critical for making reliable predictions. We expect this model will provide accurate and robust predictions of thermosphere density along various satellites in LEO.

The subsequent sections of this paper will be organized as follows: Section 2 will provide a brief description of the methodology we used, including the definition of the framework, the fundamental theorem of the deep evidential model, and the metrics we used to evaluate the accuracy of the model. Section 3 will present the databases we used for evaluation, including various databases from the CHAMP and GRACE-A, and discuss the performance of the predicted results. Then the models will be tested on the GOCE satellite, which will be used to verify its ability to handle a variety of satellite datasets. The last section will offer a comprehensive summary of our results and discuss potential applications for the future in the field of global thermosphere density prediction based on deep evidential model.

METHODOLOGY

Deep Evidential Model

Amini et al.[23] propose the deep evidential regression method by placing evidential priors over the original Gaussian likelihood function and training the neural network to infer the hyperparameters of the evidential distribution. Given a set of input variable $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^n \in \mathbb{R}^{n\times d}$ and its corresponding output $\mathbf{y} = \{y_i\}_{i=1}^n \in \mathbb{R}^n$, where n is the number of samples in the data set and d is the dimension of the input. Assume the output follows a Gaussian distribution with unknown mean and variance (μ, σ^2). The parameters are defined as $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$, and a Gaussian prior is placed on the unknown mean, and an Inverse-Gamma prior is placed on the unknown variance. These assumptions can be represented as:

$$\begin{gathered}(y_1, \ldots, y_N) \sim \mathcal{N}\left(\mu, \sigma^2\right) \\ \mu \sim \mathcal{N}\left(\gamma, \sigma^2 v^{-1}\right) \quad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta)\end{gathered} \tag{1}$$

where $\Gamma()$ is the gamma function. The hyper-parameters now can be defined as $\boldsymbol{m} = (\gamma, v, \alpha, \beta)$ and $\gamma \in \mathbb{R}$, $v > 0$, $\alpha > 1$, $\beta > 0$.

The model is trained using a novel loss function so that the network can make predictions as well as provide uncertainty estimations. The loss function combines a term that measures the distance between the predicted and true values with a term that measures the discrepancy between the predicted and actual uncertainties. The prediction, aleatoric and epistemic uncertainties can be calculated as follows:

$$Prediction : \mathbb{E}[\mu] = \gamma \tag{2}$$

$$AleatoricUncertainty : \mathbb{E}\left[\sigma^2\right] = \frac{\beta}{\alpha - 1} \tag{3}$$

$$EpistemicUncertainty : \mathrm{Var}[\mu] = \frac{\beta}{v(\alpha - 1)} \tag{4}$$

The total uncertainty is the sum of the aleatoric and epistemic uncertainty.

Prediction Framework

The framework definition now is defined as 5.

$$\lg\left(\hat{\rho}\left(t\right)\right) = f\left(\begin{array}{l} lat_{sat}(t), lon_{sat}(t), height_{sat}(t), \\ \lg\left(\rho_{JB}(t)\right), \ldots \lg\left(\rho_{JB}\left(t - D_{JB}t_s\right)\right) \\ \lg\left(\rho_{NRL}(t)\right), \ldots \lg\left(\rho_{NRL}\left(t - D_{NRL}t_s\right)\right) \\ F_{10.7}(t-1d), F_{10.7A}(t-1d) \\ Ap(t), F_{30}(t), \rho_{CHAMP}(t - t_D) \\ Dst(t - T_{Dst}), ..., Dst(t-1hr), Dst(t) \\ SymH(t - T_{SymH}), ..., SymH(t-1min), SymH(t) \\ lg(\rho_{sat}(t - t_D)) \end{array}\right) \tag{5}$$

The $\hat{\rho}$ on the left side of the equation is the predicted density from the evidential model at time t. $lat_{sat}(t), lon_{sat}(t), height_{sat}(t)$ are the latitude, longitude, and height to specify the position of the satellite at the current time t. ρ_{JB} and ρ_{NRL} are the densities estimated by the two empirical models JB2008 and NRLMSISE-00. $F_{10.7}(t-1d)$ and $F_{10.7A}(t-1d)$ refer to the daily value of $F_{10.7}$ solar flux and its 81-day averaged value with one-day lag. $1d$ is equal to 24 hours. $Ap(t)$ is derived from the 3-hour geomagnetic index K_p. $F_{30}(t)$ is the daily value of F_{30} solar index. $Dst(t)$ is the value of the magnetic activity index measuring the intensity of the globally symmetrical equatorial electrical current with one-hour resolution. $SymH(t)$ is the one-minute resolution version of the Dst index.[24] $\rho_{sat}(t)$ is the density value referred from satellite accelerometer. Through a trial and error process, we set parameters for time delays. D_{JB} and D_{NRl} are the numbers of delays in JB2008 or NRLMSISE-00, which are set as 16. T_{Dst} is the time delay of Dst. In this framework its value is 3-hr. T_{Symh} is the time delay of Symh, which is set as 15-min. There is also a time delay t_D in the CHAMP density measurement, which is set as 300 seconds. Our previous studies have demonstrated that it is necessary to provide these inputs for the model to make accurate predictions.

The data we used are all from public websites. The density derived from the empirical model JB-2008 was based on the open-source code provided by.[25] The estimated density derived from the NRLMSISE-00 model was obtained from.[26] For the geomagnetic indices $F_{10.7}$, $F_{10.7A}$, Ap, and F_{30}, we sourced the data from T.S Kelso, as referenced in.[27] To access the Dst data, we refer to,[28] while the $Symh$ data can be obtained from.[29] For the density of the satellite, in the following sections, we used the density from CHAMP and GRACE, which is referred to Mehta et al.,[30,31] and the density from GOCE is referred to.[32] Here we make the assumption that the density from the satellite accelerometer serves as the true density value as it provides the highest accuracy among all the available information and it has been widely used in the literature for the performance validation.[9,17,19,33]

Neural Network Structure

The neural network is built based on the toolbox Keras[34] in Python 3.9. We first optimize the neural network structures using the KerasTuner[35] and then make further modifications to improve the results. The neural network structure can be visualized in Figure 1.

It starts with 128 neurons connected to the inputs in the first hidden layer, followed by 256 neurons in the second layer. The third layer consists of 512 neurons, followed by 256 neurons

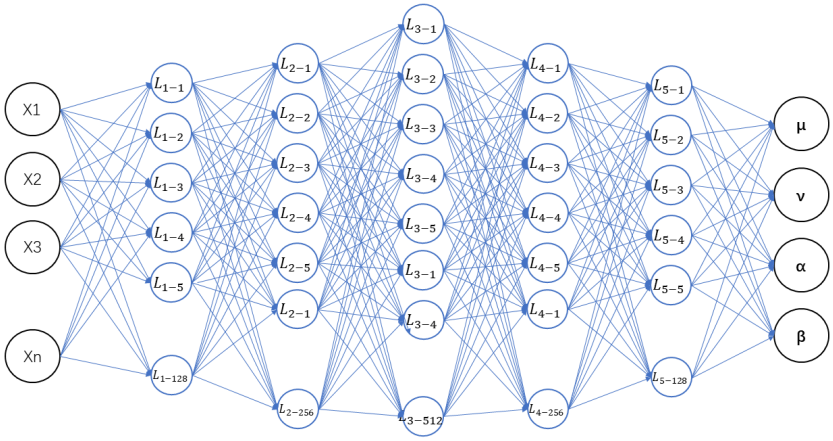

Figure 1: Neural Network Structure

in the subsequent layer. The last hidden layer, encompassing 128 neurons, is connected to four outputs representing the hyperparameters for the evidential distribution. The activation function throughout the network is rectified linear unit (Relu). During the training process, the model is trained for 500 epochs. To prevent overfitting and enhance efficiency, early stopping is applied when the accuracy of the validation section is degraded after 15 steps. These configuration settings optimize the performance and training of the neural network for the given task. We standardized the data before training.

Performance Metrics

To evaluate the performance of the proposed model, four metrics are used in this paper to analyze the results. To assess the accuracy of the predictions, we use the Pearson correlation coefficient (R) and the Root Mean Squared Error (RMSE).

The definition of R and RMSE can be mathematically expressed as Eqs. 6 and 7:

$$R = \frac{\sum_{i=1}^{n} (\rho_i - \bar{\rho}) \left(\hat{\rho}_i - \bar{\hat{\rho}}\right)}{(n-1)\sigma_\rho \sigma_{\hat{\rho}}} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{\rho}_i - \rho_i)^2} \tag{7}$$

where ρ_i and $\hat{\rho}_i$ are the true density and predicted density. $\bar{\rho}$ represents the mean value of the density. σ_ρ and $\sigma_{\hat{\rho}}$ are the standard deviations of the truth and the predictions, and n is the size of the data that are used for evaluation. A good performance shall have an R close to one and an RMSE as small as possible.

To study the uncertainty prediction performance, we calculate the coverage rate of 2σ area (Cov Rate) to evaluate the quality of the uncertainty. The coverage rate is defined as Eq. 8:

$$\text{Cov Rate } = \frac{k}{n} \times 100\% \tag{8}$$

where k is the number of the true density that is within the 2σ uncertainty boundaries estimated by the evidential model. A good performance shall have a Coverage Rate close to 100%.

We also evaluate the confidential level and calculate the Mean Absolute Calibration Error (MACE) to evaluate the reliability of the uncertainty. Calibration measures a model's predicted probabilities, and a well-calibrated model is one in which the predicted probabilities are reliable and trustworthy. It is important for a model to be well-calibrated in order to make effective use of predicted probabilities in further analysis.

The confidence interval range is defined as $CL = [5\%, 10\%, ..., 95\%, 99\%]$. The corresponding coefficients defining the uncertainty bounds are then given as Eq. 9, where erf is the error function.

$$\zeta[k] = \sqrt{2}\,\mathrm{erf}^{-1}(C[k]/100) \tag{9}$$

To evaluate the reliability of the uncertainty, we calculate the mean absolute calibration error (MACE), defined as Eq. 10.

$$MACE = \frac{1}{n_C}\sum_{k=1}^{n_C}|C[k] - P[k]| \tag{10}$$

The MACE calculates the average difference between the predicted probability and the actual frequency in the test set, allowing us to assess the overall reliability of the model's prediction. The lower values of the MACE indicate better-calibrated models. We apply a scalar factor[36] to the standard deviation of the predicted value to help improve the overall calibration of the model. The scalar factor is calculated by Eq. 11.

$$s = \sqrt{\frac{1}{n_v}\sum_{i=1}^{n_v}\left[\frac{\|\mathbf{y_i} - \hat{\mu}_\mathbf{i}\|^2}{\hat{\sigma}_i^2}\right]} \tag{11}$$

where n_v is the number of validation data, $\mathbf{y_i}$ is the true data, $\hat{\mu}_\mathbf{i}$ is the predicted mean value, and $\hat{\sigma}_i^2$ is the predicted standard deviation for the i^{th} validation sample.

CASE STUDIES AND RESULTS

Train and Test on CHAMP and GRACE

The databases used in this section are from two different satellites: CHAMP and GRACE-A.

We assume we have access to long periods of data from CHAMP, and the training database from CHAMP in this section is selected from 01/01/2002 to 07/31/2005. For GRACE-A, on the other hand, we assume our access is limited to a specific period, and in this section, we can only access the data from 07/01/2004 to 12/31/2004. In the subsequent experiments, we aim to train the models using different databases. Each model represents a specific combination of datasets from the CHAMP and GRACE-A satellites, defined by distinct periods. Model-1 uses the data only from CHAMP, covering three and a half years from 2002 to 2005. Model-2 only uses data from GRACE-A and covers only one month in 2004. Model-3 only uses data from GRACE-A too, but the period is longer than Model-2, which covers half of the year in 2004. The size of the training data in Model-3 makes the local predicted reasonable. Model-4 contains the large number of data from CHAMP as

Model-1 and the same length of the GRACE-A data as Model-2. We can compare the results of Model-1, Model-2 and Model-4 to see the effect of the additional GRACE-A data. In Model-5 the database contains the large number of CHAMP data, and one day from GRACE-A. To clarify, we define the models with different training databases and summarize them as Table 1.

Table 1: Definitions of Models

Model	Database
Model-1	01/01/2002 to 07/31/2005 in CHAMP.
Model-2	12/01/2004 to 12/31/2004 in GRACE-A.
Model-3	07/01/2004 to 12/31/2004 in GRACE-A
Model-4	01/01/2002 to 07/31/2005 in CHAMP, and 12/01/2004 to 12/31/2004 in GRACE-A.
Model-5	01/01/2002 to 07/31/2005 in CHAMP, and 12/01/2004 in GRACE-A

The altitudes of the used periods for the two satellites are plotted in Figure 2. The blue section corresponds to the altitude of CHAMP from 01/01/2002 to 07/31/2005. The boundaries of this blue section represent the range of altitudes the CHAMP covers, and the bolded line indicates the daily mean value of the altitude. Similarly, the red section is the altitude of GRACE-A from 01/01/2002 to 07/31/2005. Notice only short periods of Grace data will be available for training. The altitudes of the two satellites exhibit significant differences, with no overlapping section between them. The dissimilarity in altitude further underscores the distinct characteristics and orbits of CHAMP and GRACE-A.

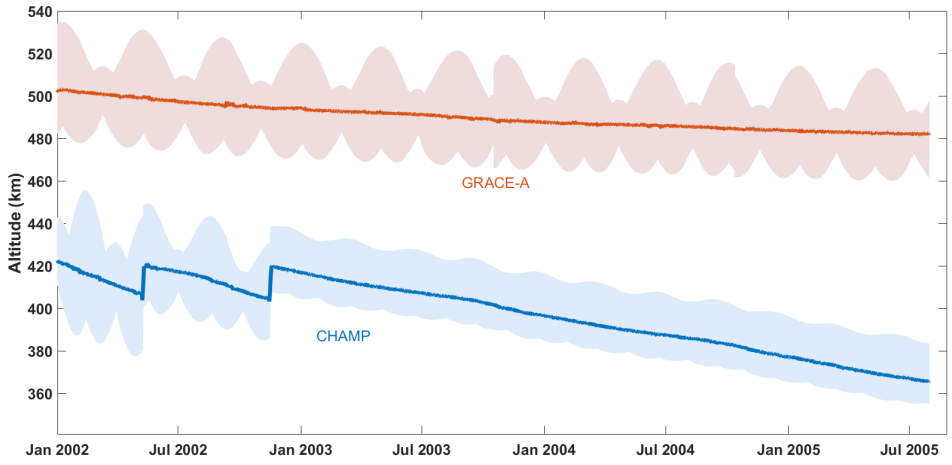

Figure 2: Altitudes of satellites with bold lines representing the satellites' mean altitude: CHAMP and GRACE-A

We will evaluate the model performances on several test cases based on different models. The test period of the test cases is stated in Table 2.

Test-1 is the Halloween storm event, containing the second biggest storm in 2003. This period for CHAMP is included in the training range.

Test-2 is from 08/31/2005 to 09/01/2005. The test section is not included in the training data, but the test section is in a further future than the training section. There is a storm that happened during

Table 2: Definitions of Test Cases

Test case	Test Period
Test-1	10/27/2003 to 11/03/2003
Test-2	08/31/2005 to 09/01/2005

this period. We will evaluate the performance of both CHAMP and GRACE-A for the test cases.

We note that the results in the following sections are obtained by averaging the performance from 10 random seeds from each model.

Test-1 The predicted results from 10/27/2003 - 11/03-2003 from these models are presented in Table 3.

Table 3: Test on 10/27/2003 - 11/03/2003

CHAMP	Model-1	Model-2	Model-3	Model-4	Model-5
R	0.9861	0.8620	0.8821	0.9859	0.9860
RMSE($\times 10^{-12} kg/m^3$)	0.4874	1.7872	1.3741	0.4907	0.4880
CovRate	0.9882	0.0037	0.0951	0.9841	0.9846
MACE	0.0309	0.2848	0.2601	0.0370	0.0376
GRACE	Model-1	Model-2	Model-3	Model-4	Model-5
R	0.8642	0.8394	0.8401	0.8672	0.8242
RMSE ($\times 10^{-12} kg/m^3$)	0.3088	0.6296	0.3696	0.2795	0.2954
CovRate	0.4027	0.1981	0.8671	0.8736	0.7558
MACE	0.1025	0.3735	0.2440	0.1476	0.2545

For the test on CHAMP, we can see that the results on Model-1, Model-4, and Model-5 have high accuracy and reliable uncertainty estimations with a large amount of CHAMP data in the training databases. They show very high R values indicating a strong relationship between the predicted values and the actual data. They also show small RMSE values. The coverage rate values for the three models are all beyond 98%, which means most of the truth values can be covered within the uncertainty boundaries. The MACE values are very close to zero. The two metrics indicate the uncertainty estimations are reasonable and reliable. We can also see that the performances of Model-4 and Model-5 are slightly worse than Model-1. The reason is that the data distribution GRACE-A covered by Model-4 and Model-5 significantly differs from the distribution used in Model-1.

The predicted results of CHAMP from Model-2 and Model-3 are not reasonable because the training database does not cover any information from CHAMP. The accuracy of the predicted results from Model-2 and Model-3 is much worse than the results from the other model models. The R value reduced from 0.98 to 0.86, while the RMSE value increased from 0.4 to 1.7 ($\times 10^{-12} kg/m^3$). The coverage rate from the models is very low, which indicates that the model cannot capture the full range of variations in the data.

For the test on GRACE, Model-4 shows the best results as it has the highest R and coverage rate values and the smallest RMSE and MACE values. Model-1's training database does not contain any information about the GRACE, so the uncertainty estimations are very bad. The coverage rate in Model-1 is only 40%. Model-2, which uses one-month data from GRACE-A, shows the lowest R and Coverage Rate and highest RMSE and MACE values, indicating the model cannot learn the data

feature well with such a small size of data. When the training period from GRACE-A is extended, the performance from Model-3 is enhanced much better than Model-2, with a higher Coverage Rate and more moderate RMSE and MACE values, suggesting that it can capture most of the data features for GRACE. As for the performance in Model-5, which contains the large size of CHAMP data and one-day GRACE-A data, the predicted accuracy is very close to the results in Model-1, but the uncertainty estimation from Model-5 is much better than Model-1. Comparing the results from Model-3 with Model-5, Model-5 shows a smaller RMSE value. But Model-5 cannot give a good uncertainty estimation with the small size of GRACE-A data when testing on GRACE-A, so the coverage rate and MACE value are worse than those in Model-3. Overall, Model-4 appears to be the top-performing model on the GRACE dataset, demonstrating high accuracy in predicted results and reasonable uncertainty estimations.

We plot the residuals of the predictions on GRACE-A in Figure 3 with Model-1, Model-3, Model-4, and Model-5, ignoring the performance of Model-2 because of the bad results. The blue section represents the results from Model-1, the red section is the results from Model-3, the yellow section corresponds to Model-4, and the purple one from Model-5. The residual distribution along the test times is presented on the left-hand side. On the right-hand side is the histogram distribution of the residuals, and the approximated Gaussian distributions of the residuals. From Figure 3 we can see the largest error is from Model-3. The mean and standard deviation values from Model-1 and Model-5 are close. Model-4 shows a mean residual value closest to zero, indicating its predictions are more accurate.

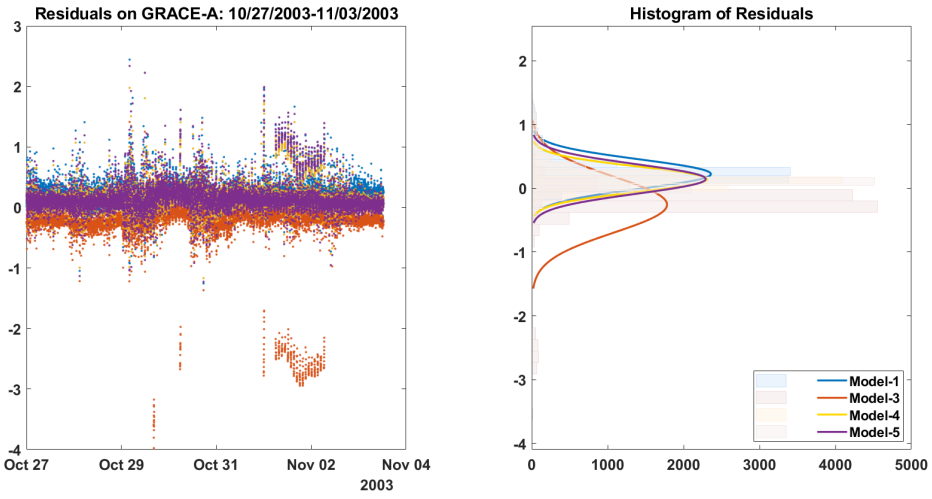

Figure 3: Residuals on GRACE-A: 10/27/2003-11/03/2003

Test-2 The test period in Test-2 is from 08/31/2005 to 09/01/2005. It is one month after the training range. The predicted results for the CHAMP and GRACE-A are presented in Table 4.

For the predicted results on CHAMP, Model-1, Model-4 and Model-5 demonstrate high correlation coefficients (R) close to 0.96. The RMSE values in the two models are close to 0.31 ($\times 10^{-12} kg/m^3$), suggesting the predicted results are accurate. The coverage rate on the two models is beyond 98%, and MACE values are quite low, indicating the estimated uncertainty is quality and reliable. Model-4 and Model-5 show a slightly worse tendency than Model-1 for adding more data from GRACE-A, while with smaller data from GRACE-A, the predicted performance from Model-

Table 4: Test on 08/31/2005 - 09/01/2005

Champ	Model-1	Model-2	Model-3	Model-4	Mdoel-5
R	0.9579	0.8679	0.8743	0.9571	0.9775
RMSE($\times 10^{-12} kg/m^3$)	0.3112	1.0845	0.5294	0.3162	0.3143
CovRate	0.9854	0.0028	0.1458	0.9843	0.9861
MACE	0.0330	0.3357	0.1351	0.0446	0.0363
GRACE	Model-1	Model-2	Model-3	Model-4	Model-5
R	0.9259	0.8706	0.8761	0.9774	0.8759
RMSE(($\times 10^{-13} kg/m^3$)	0.7372	0.6520	0.5047	0.4840	0.5385
CovRate	0.7159	0.3662	0.9216	0.9409	0.9067
MACE	0.1074	0.1573	0.0497	0.0491	0.0726

5 is better than Model-4. On the other hand, for Model-2 and Model-3 with information only from GRACE-A, the performance becomes much worse. The worst performance is from Model-2, with the R value at 0.86, while the RMSE value increased to 1.0845 $\times 10^{-12}$. Both Model-2 and Model-3 show lower coverage rates with larger MACE values than the other two, indicating that the models reproduce the predictions with less accurate predictions and unreasonable uncertainty estimation.

As for the results of GRACE-A, Model-1 shows the largest RMSE value among all the models since the training database does not contain any information about GRACE-A. For models trained only with GRACE-A, The performance from Model-3 is more accurate than the results from Model-2 by having a longer training period. Model-3 also gives a reasonable uncertainty estimation because the coverage rate is improved from 36% to 92% compared with Model-2. Model-5 in this case, does not show the advantages of Model-3 with only one-day GRACE-A data in the training database. But the RMSE value and the uncertainty estimations are better than Model-1 and Model-2. Model-4 shows the best performances with the largest R and coverage rate values and the smallest RMSE and MACE values.

The residual features of each model in this test case is plotted in Figure 4. The distribution from Model-1, Model-4 and Model-5 are close, while the residuals from Model-3 show a larger bias, indicating consistency between the distribution patterns and error magnitudes.

The experiments demonstrate that combining CHAMP and limited GRACE data significantly improves predictions compared to using only CHAMP or GRACE data alone. Despite limitations, combining a large number of CHAMP data with shorter-period GRACE data can provide valuable insights that refine the model's predictive power. This approach shows great potential in constructing a robust global thermospheric density prediction model for diverse satellite missions.

Test on GOCE

In the previous section, our experiments use data from the CHAMP and GRACE satellites, and the test periods are also drawn from these two satellites. In this section, we extend our testing of the proposed model to include data from the GOCE satellite to prove its effectiveness as a global prediction framework.

Unlike the CHAMP and GRACE satellites, there is no information about GOCE in the training

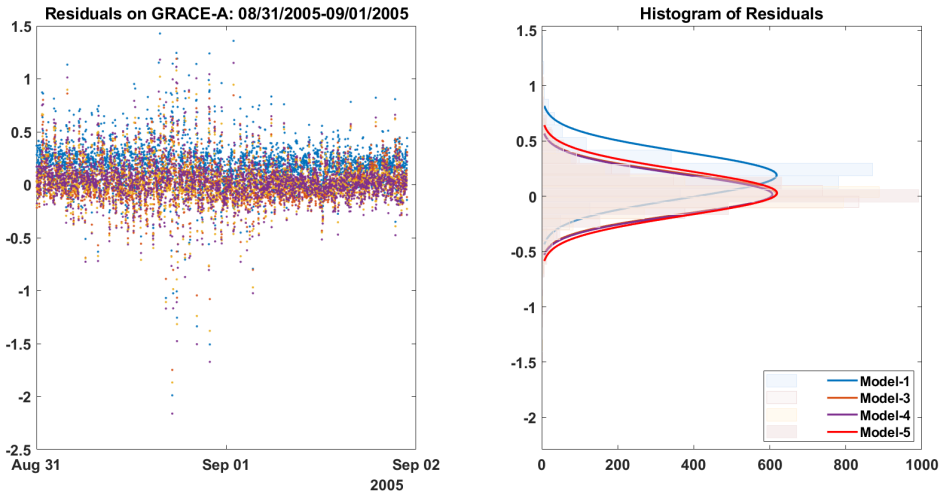

Figure 4: Residuals on GRACE-A: 08/31/2005-09/01/2005

database. The test period in GOCE is from 01/01/2010 to 01/31/2010, which is far further than the training date. The altitude of the GOCE is also quite different from the GRACE and CHAMP.

The well-trained model, as studied in the previous section, will also be used for this test case. Considering that the test information from GOCE has not been included in the training dataset, the test cases are expected to be more challenging for prediction. To enhance the model's performance, we have extended the training database and introduced as Model-6. Model-6 is defined as training from 01/01/2002-07/31/2005 from both CHAMP and GRACE. The altitude of the periods from the three satellites is plotted in Figure 5.

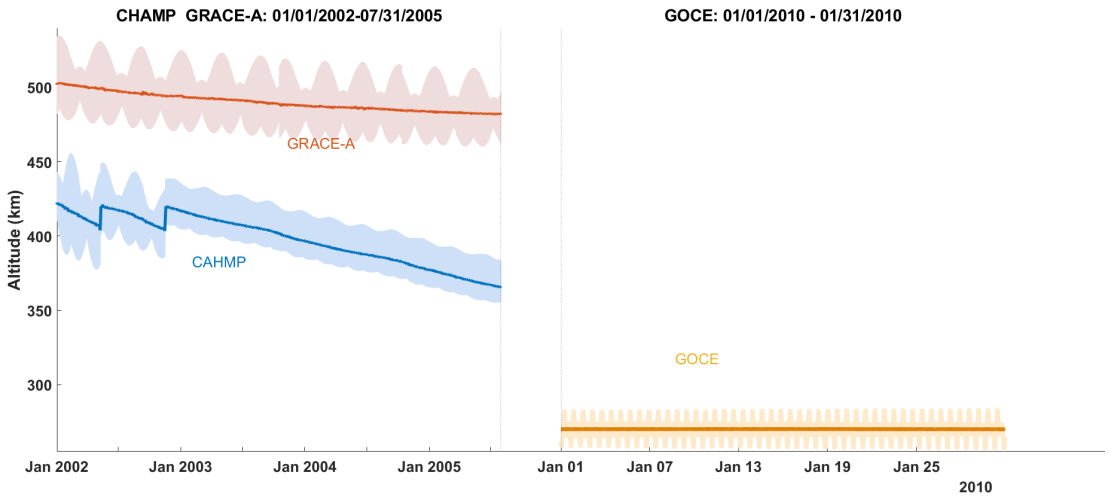

Figure 5: Altitudes of satellites with bold lines representing the satellites' mean altitude: CHAMP, GRACE-A and GOCE. Notice the time scale for GOCE is different from the other two satellites.

From Figure 5 we can see the altitude of GOCE during 01/01/2010 - 01/31/2010 is lower than the other two satellites from 01/01/2002 to 07/31/2005. It is worth noting that the altitude of GOCE is

not included in the training dataset, indicating that the model has not been exposed to this specific satellite's altitude profile during its learning process. The results from the models are presented in Table 5.

Table 5: Test on GOCE 01/01/2010-01/31/2010

Model	Model-1	Model-2	Model-3	Model-4	Model-5	Model-6
R	0.8569	0.3627	0.7681	0.8730	0.8708	0.9646
RMSE	1.4369	5.5709	3.4754	1.0056	1.2277	0.9635
CovRate	0.7448	0.0001	0.0031	0.8538	0.7635	0.9098
MACE	0.0915	0.5940	0.3090	0.0727	0.0757	0.0587

Accordingly, Model-1. Model-4 and Model-5 provide satisfactory results based on the R values exceeding 0.85 and the coverage rates exceeding 74%. Model-2 and Model-3 cannot learn the data feature well so the results from the two models are unreasonable. Model-4, Model-5, and Model-6 show better performance than Model-1, with higher R value and coverage rate, and smaller RMSE and MACE value. Even Model-5 contains the GRACE-A data last only one day, the accuracy of the predictions is still better than Model-1. Model-6 in this case shows the best results with the extended training base. When the training database contains both CHAMP and GRACE-A data, even if the GRACE-A data last only one day, the predicted results are better than the model with training data from one satellite. These studies have demonstrated that despite the distinctive density distribution derived from the GOCE satellite during the study, the models can produce reasonable predictions well supported by the data.

The residuals of predictions on GOCE are plotted in Figure 6. The blue section represents the results from Model-1, the red section corresponds to Model-2, the yellow section displays the results from Model-3, the purple section is from Model-4, the green section shows the residuals from Model-5, and the light blue section is for Model-6. The left-hand side of the figure illustrates the residual distribution along the test time. On the right-hand side, it shows the histogram of the residuals and the approximated Gaussian distribution curve, providing a clearer perspective on their characteristics. The predictions from Model-2 and Model-3 show larger errors than the other models, which correspond to the lower accuracy performance from Table 5. Model-1, Model-4, and Model-5 predictions are slightly smaller than the truth value since the mean of the residuals is smaller than zero. The mean value of the residuals from Model-1 and Model-5 are close. The mean value of the residuals from Model-4 is closer to zero compared with Model-1 and Model-5. As for Model-6, We can see the center of the purple section is the closest to zero, indicating a smaller bias in the predictions. Moreover, the purple section shows a smaller standard deviation, which indicates the predictions from Model-6 is more accurate than the others.

The results of testing on GOCE show that our proposed model can be applied to a wider range of satellite density predictions, even if the satellites are located in different positions. Given the distinct characteristics of GOCE data and its absence from our training database, this analysis is a robust assessment of the model's generalization capabilities and its potential as a truly global thermospheric density prediction model.

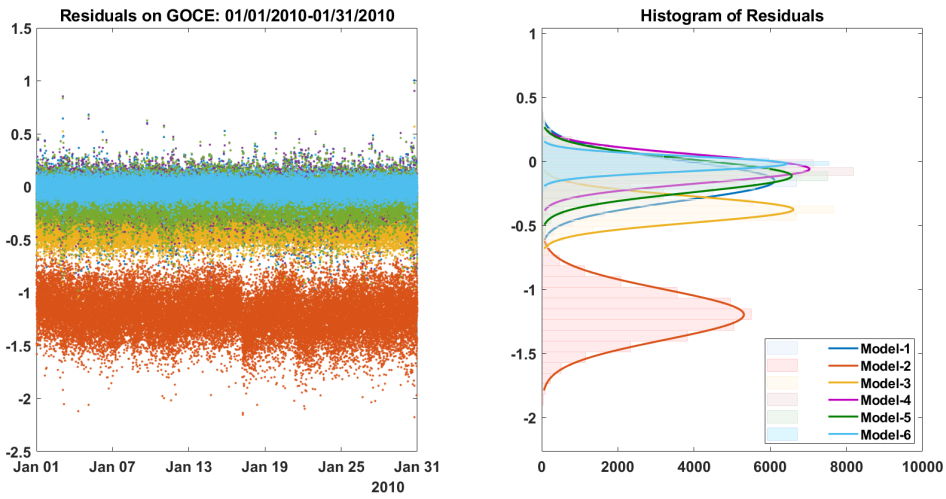

Figure 6: Residuals on GOCE Predictions

CONCLUSION

This paper proposes a evidential model-based framework for predicting global thermospheric density across various satellites, such as CHAMP, GRACE, and GOCE. By using a large number of CHAMP data and a limited GRACE dataset, we demonstrate the model's effectiveness in achieving high accuracy and providing reliable uncertainty estimation results. Our findings highlight the potential of this combined database approach in delivering robust predictions for thermospheric density on different satellites even the new satellite information is not covered in the training database.

In the first section of the experiments, different training databases are used to evaluate different test cases. Our results demonstrate that model with hybrid training database significantly improved performance compared to the model using data from only one satellite. Specifically, the combination of the large number of CHAMP data and the valuable insights from the limited GRACE data enhances the model's predictive capabilities, leading to more accurate and reliable results.

Furthermore, when extending the evaluation to test on the GOCE satellite, whose test period is out of the training range and altitude is also different from the CHAMP and GRACE satellites, the predicted performance from the model with the combined databases is satisfying. The predicted R values exceed 0.87, and coverage rates are beyond 85%, indicating their capacity to adapt and provide reliable predictions across diverse satellite missions.

In conclusion, the proposed strategy shows its potential for predicting the thermospheric density of various satellite missions with high accuracy and reliable uncertainty estimations, even when faced with distinct density distributions from satellites with varying altitudes. These results contribute to advancing space environment research and developing accurate density prediction models for satellite operations.

REFERENCES

[1] "The Thermosphere," https://scied.ucar.edu/learning-zone/atmosphere/thermosphere.

[2] J. Picone, A. Hedin, D. P. Drob, and A. Aikin, "NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues," *Journal of Geophysical Research: Space Physics*, Vol. 107, No. A12, 2002, pp. SIA–15.

[3] B. Bowman, W. K. Tobiska, F. Marcos, C. Huang, C. Lin, and W. Burke, "A new empirical thermospheric density model JB2008 using new solar and geomagnetic indices," *AIAA/AAS astrodynamics specialist conference and exhibit*, 2008, p. 6438.

[4] "The Thermospheric General Circulation Models (TGCM's)," http://www.hao.ucar.edu/modeling/tgcm.

[5] "Global Ionosphere Thermosphere Model (GITM)," https://ccmc.gsfc.nasa.gov/models/modelinfo.php?model=GITM.

[6] J. Emmert, "A long-term data set of globally averaged thermospheric total mass density," *Journal of Geophysical Research: Space Physics*, Vol. 114, No. A6, 2009.

[7] D. Weimer, B. Bowman, E. Sutton, and W. Tobiska, "Predicting global average thermospheric temperature changes resulting from auroral heating," *Journal of Geophysical Research: Space Physics*, Vol. 116, No. A1, 2011.

[8] S. Elvidge, H. C. Godinez, and M. J. Angling, "Improved forecasting of thermospheric densities using multi-model ensembles," *Geoscientific Model Development*, Vol. 9, No. 6, 2016, pp. 2279–2292.

[9] C. Xiong, H. Lühr, M. Schmidt, M. Bloßfeld, and S. Rudenko, "An empirical model of the thermospheric mass density derived from CHAMP satellite," *Annales Geophysicae*, Vol. 36, Copernicus Publications Göttingen, Germany, 2018, pp. 1141–1152.

[10] D. R. Weimer, M. Mlynczak, J. Emmert, E. Doornbos, E. Sutton, and L. Hunt, "Correlations between the thermosphere's semiannual density variations and infrared emissions measured with the SABER instrument," *Journal of Geophysical Research: Space Physics*, Vol. 123, No. 10, 2018, pp. 8850–8864.

[11] P. M. Mehta, R. Linares, and E. K. Sutton, "Data-driven inference of thermosphere composition during solar minimum conditions," *Space Weather*, Vol. 17, No. 9, 2019, pp. 1364–1379.

[12] D. J. Gondelach and R. Linares, "Real-time thermospheric density estimation via two-line element data assimilation," *Space Weather*, Vol. 18, No. 2, 2020, p. e2019SW002356.

[13] P. M. Mehta and R. Linares, "Real-time thermospheric density estimation from satellite position measurements," *Journal of Guidance, Control, and Dynamics*, Vol. 43, No. 9, 2020, pp. 1656–1670.

[14] D. J. Gondelach and R. Linares, "Real-Time Thermospheric Density Estimation via Radar and GPS Tracking Data Assimilation," *Space Weather*, Vol. 19, No. 4, 2021, p. e2020SW002620.

[15] R. J. Licata and P. M. Mehta, "Reduced Order Probabilistic Emulation for Physics-Based Thermosphere Models," *Space Weather*, Vol. 21, No. 5, 2023, p. e2022SW003345.

[16] V. Nateghi, "Machine learning-based thermospheric density modelling and estimation for space operations," 2021.

[17] S. Bonasera, G. Acciarini, J. A. Pérez-Hernández, B. Benson, E. Brown, E. Sutton, M. K. Jah, C. Bridges, and A. G. Baydin, "Dropout and Ensemble Networks for Thermospheric Density Uncertainty Estimation,"

[18] R. J. Licata, P. M. Mehta, W. K. Tobiska, and S. Huzurbazar, "Machine-Learned HASDM Thermospheric Mass Density Model With Uncertainty Quantification," *Space Weather*, Vol. 20, No. 4, 2022, p. e2021SW002915.

[19] T. Gao, H. Peng, and X. Bai, "Calibration of atmospheric density model based on Gaussian Processes," *Acta Astronautica*, Vol. 168, 2020, pp. 273–281.

[20] Y. Wang, H. Peng, X. Bai, J. T. Wang, and H. Wang, "Advance Thermospheric Density Predictions through Forecasting Geomagnetic and Solar Indices Based on Gaussian Processes," *AAS/AIAA Astrodynamics Specialist Conference*, 2021.

[21] Y. Wang and X. Bai, "Comparison of Gaussian processes and Neural Networks for thermospheric density predictions during quiet time and geomagnetic storms," *2022 AAS/AIAA Astrodynamics Specialist Conference*, 2022.

[22] Y. Wang and X. Bai, "Thermospheric density predictions during quiet time and geomagnetic storm using a deep evidential model-based framework," *Acta Astronautica*, 2023.

[23] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *Advances in Neural Information Processing Systems*, Vol. 33, 2020, pp. 14927–14937.

[24] J. A. Wanliss and K. M. Showalter, "High-resolution global storm index: Dst versus SYM-H," *Journal of Geophysical Research: Space Physics*, Vol. 111, No. A2, 2006.

[25] M. Mahooti, "Jacchia-Bowman Atmospheric Density Model," https://www.mathworks.com/matlabcentral/fileexchange/56163-jacchia-bowman-atmospheric-density-model.

[26] M. Mahooti, "NRLMSISE-00 Atmosphere Model," https://www.mathworks.com/matlabcentral/fileexchange/56253-nrlmsise-00-atmosphere-model.

[27] T. S. Kelso, "Space Weather Data Documentation," https://celestrak.org/SpaceData/SpaceWx-format.php.

[28] K.-O. Cho, "Geomagnetic equatorial DST index home page," https://wdc.kugi.kyoto-u.ac.jp/dstdir/.

[29] "ASY/SYM INDICES," https://isgi.unistra.fr/indices_asy.php.

[30] P. M. Mehta, A. C. Walker, E. K. Sutton, and H. C. Godinez, "New density estimates derived using accelerometers on board the CHAMP and GRACE satellites," *Space Weather*, Vol. 15, No. 4, 2017, pp. 558–576.

[31] "CHAMP and GRACE Density Data Sets," https://drive.google.com/drive/folders/0BwtX8XEH-aEueHJiU1htLVo0cms?resourcekey=0-byxPMLbZSVC5Blxb_Rfl9g.

[32] "ESA GOCE Online Dissemination: GOCE Thermosphere Data," https://goce-ds.eo.esa.int/oads/access/collection/GOCE_Thermosphere_Data/tree?p0=TDC_GOC_2_.

[33] D. Pérez, B. Wohlberg, T. A. Lovell, M. Shoemaker, and R. Bevilacqua, "Orbit-centered atmospheric density prediction using artificial neural networks," *Acta Astronautica*, Vol. 98, 2014, pp. 9–23.

[34] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[35] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, *et al.*, "Keras Tuner," https://github.com/keras-team/keras-tuner, 2019.

[36] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, "Recalibration of aleatoric and epistemic regression uncertainty in medical imaging," *arXiv preprint arXiv:2104.12376*, 2021.