Acta Astronautica 211 (2023) 316–325

Contents lists available at ScienceDirect

Acta Astronautica

journal homepage: www.elsevier.com/locate/actaastro

Research paper

Thermospheric density predictions during quiet time and geomagnetic storm using a deep evidential model-based framework

Yiran Wang[1], Xiaoli Bai [*,2]

Department of Mechanical and Aerospace Engineering, Rutgers, The State University of New Jersey, NJ 08854, USA

ARTICLE INFO

Keywords:
Thermospheric density prediction
Deep evidential model
Uncertainty prediction

ABSTRACT

Knowledge of the thermospheric density is essential for calculating the drag in low Earth orbit satellites. Existing models struggle to predict density accurately. In this paper, we propose thermospheric density prediction using a deep evidential model-based framework that incorporates empirical models, accelerometer-inferred density from the CHAMP satellite, and geomagnetic and solar indices. The framework is investigated on both quiet and storm conditions. Our results demonstrate that the proposed model can predict the thermospheric density with high accuracy and reliable uncertainty in both quiet and storm times. The predicted results from the evidential model are advantageous over the Gaussian Processes (GPs) model in our previous studies. Furthermore, the proposed model can also provide insightful aleatoric and epistemic uncertainties.

1. Introduction

Atmospheric drag is the dominant perturbation force and is the most difficult to predict for orbit propagation of low Earth orbit satellites [1]. It is influenced by the drag coefficient, thermospheric density, the area of the object facing the fluid, and the relative velocity of the satellite. Knowledge of thermospheric density is essential for calculating the drag of LEO satellites and the accuracy of the thermospheric density will affect the prediction of the satellite trajectory. The thermospheric density also plays a role when space debris reenter the Earth's atmosphere [2]. Accurate information on the thermospheric density leads to more precise predictions of re-entry time and location, thereby reducing the potential harm to people and property on the ground.

The study of thermospheric density prediction in space weather, driven by the need for accuracy and reliability in predicting the atmospheric density that affects the performance and safety of the spacecraft, has evolved significantly over the years. Early efforts to predict the density focus on using physical laws and observations to develop empirical models, such as Naval Research Laboratory Mass Spectrometer and Incoherent Scatter Radar Extended (NRLMSISE) [3] and Jacchia-Bowman (JB) [4] models. Earth's atmosphere is a complex system affected by various factors, such as Solar flares, high-speed solar winds, etc. Besides, the location and the corresponding temperature also affect the atmospheric density [5]. Empirical physical-based models struggle to capture the full complexity of the atmosphere, and the accuracy of their predictions is limited. The empirical models are likely to generate huge errors during periods of high solar and geomagnetic activities, and the uncertainties could reach beyond 100% under some extreme conditions [6]. On the other hand, physics-based models such as the Thermospheric General Circulation Models [7] and Global Ionosphere Thermosphere Model [8] can overcome the limitation of the empirical models and are critical for advancing our knowledge of atmospheric physics. However, they currently suffer from uncertain input and boundary conditions and have failed to outperform the empirical models.

Studies indicate that the thermospheric density changes during geomagnetic storms [9–11], which in turn lead to the fluctuation of drag on satellites [12]. The magnetic storm is influenced by many factors, mainly including solar winds, seasonal and solar cycle variations [13], coronal mass ejections, Earth's magnetic field, geomagnetic activity, and other determinations [14]. The magnetic activity index Disturbance Storm Time (Dst) is commonly used to classify whether or not a storm has occurred band to define the duration of a storm. Dst is also used to distinguish between quiet and strongly disturbed geomagnetic conditions. Commonly, a minor storm is defined when Dst is between −30 nT to −50 nT; a moderate storm is defined when Dst is between −50 nT to −100 nT; an intense storm is defined when Dst is smaller than −100 nT, and the great storm is defined when Dst is smaller than −250 nT [15]. In this paper, we will study predicting thermospheric density in different space weather conditions. Two criteria have been

* Corresponding author.
E-mail addresses: yw619@rutgers.edu (Y. Wang), xiaoli.bai@rutgers.edu (X. Bai).
[1] Graduate Student.
[2] Associate Professor.

https://doi.org/10.1016/j.actaastro.2023.06.023
Received 4 March 2023; Received in revised form 14 May 2023; Accepted 13 June 2023
Available online 19 June 2023

defined for classification. Quiet time is defined as the period when the minimum Dst is larger than −50 nT, and storm time is defined as the period when the minimum Dst is smaller than −100 nT.

Over the past years, researchers have developed new methods to predict thermospheric density during either quiet or storm time. Zhou et al. [16] introduce a multiple linear regression analysis with proper time shifts to study the atmospheric density during the storm time. Liu et al. [17] investigate the thermospheric density of the merging electric field during magnetic storms. Perez et al. [18] use artificial neural networks to predict the density value. Xiong et al. [19] establish an empirical model named CH-Therm-2018 using nine years of accelerometer measurements from the CHAMP satellite with seven key parameters, including height, solar flux index, day of the year, local magnetic time, geographic latitude, longitude, and the magnetic activities represented by the solar wind merging electric field. Oliveira et al. [2] investigate the effects of satellite drag at low earth orbit during magnetic storms. They use the GRACE and CHAMP data to estimate drag from historical events. Bonasera et al. [20] 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. Gondelach et al. [21] develop a dynamic reduced-order density model to estimate the density using the TLE data, and a Kalman filter to quantify the uncertainty in the estimates. Richard et al. [22] build the model based on the Principal Component Analysis (PCA) method and test the density along the satellite orbit from 2002 to 2010. It is worth noticing that machine learning methods have recently become increasingly popular in predicting thermospheric density [23]. Compared with traditional empirical models, these methods could be highly effective, providing better accuracy and reliability.

Our earlier studies [24,25] have proposed a density estimation framework based on Gaussian Processes (GPs) that integrates information from empirical models, environment conditions, and satellite measurement data. We have demonstrated that the framework can be used to predict thermospheric density during quiet time. In a recent study [26], the model definition has been improved, and both GPs and deep neural networks have been used to predict the density during quiet and storm times. This study has shown advanced results than the work from Perez [18] during quiet time and from Liu et al. [17] during storm time in 2004.

This paper aims to incorporate uncertainty estimation into deep neural networks to investigate the accuracy and precision of the predictions. At the same time, we hope to reveal the underlying sources of uncertainty [27,28]. Quantifying model uncertainty is essential as it provides a measure of confidence in the predictions made by the model. In machine learning, there are two important types of uncertainty: aleatoric and epistemic. Aleatoric uncertainty refers to the inherent randomness or variability in the process, which can result in uncertainty even with perfect knowledge of the underlying physics or data. It could arise from measurement noise or natural variability in the process being studied. Epistemic uncertainty, on the other hand, arises from incomplete knowledge or understanding of a system. This can be due to limited or inaccurate data, incomplete models, or other sources of uncertainty. Different from aleatoric uncertainty, epistemic uncertainty can be reduced by conducting additional research or obtaining more data. Together aleatoric and epistemic uncertainty quantifies the sources of uncertainty and can provide guidance on the potential to improve the accuracy and reliability of models.

The deep evidential model (DEM), proposed by Amini et al. [29], is a machine learning method that can estimate both the aleatoric and epistemic uncertainties in training a deep neural network. In this paper, we explore using DEM to build a data-driven framework for thermospheric density prediction.

This paper makes four contributions. First, to the best of our knowledge, this is the first time that a deep evidential learning method has been used for thermospheric density prediction. This innovative use of the deep evidential model provides a new and improved method for thermospheric density prediction compared to the GPs model in our previous studies [24–26]. Second, the proposed model accurately predicts thermospheric density in both quiet and storm times. The predicted results demonstrate that the evidential model built using the same neural network structure is stable and robust in both quiet and storm times. This makes it useful in a wide range of space weather conditions. Third, the evidential model generates high-quality uncertainty prediction for both quiet and storm times. Most of the truth data are within the uncertainty boundaries, and the predictions have high reliability according to the confidential level. Fourth, the proposed evidential model can provide information about both aleatoric and epistemic uncertainties. During the quiet time, the mean value of the aleatoric is close to a constant value. When the storm happens, the density is more unpredictable and the process becomes more random than in the quiet period. It is always observed that the aleatoric uncertainty increases as the storm begins and returns to the constant value as the storm subsides. Furthermore, the total uncertainty value always increases during the storm period.

The rest of this paper is organized as follows. Section 2 describes the methodology in detail, including the basic algorithm of the deep evidential method and the proposed neural network model. We also introduce the model definition and the metrics used to evaluate model performance. Section 3 describes the database we used for evaluation including both the quiet and storm times, and discusses the performance of the predictions. Conclusions are presented in the last section.

2. Methodology

2.1. Deep evidential model

Amini et al. [29] 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 $\theta = (\mu, \sigma^2)$, 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:

$$\begin{aligned} (y_1, \ldots, y_N) &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu \sim \mathcal{N}(\gamma, \sigma^2 \upsilon^{-1}) \quad \sigma^2 &\sim \Gamma^{-1}(\alpha, \beta) \end{aligned} \tag{1}$$

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

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:

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

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

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

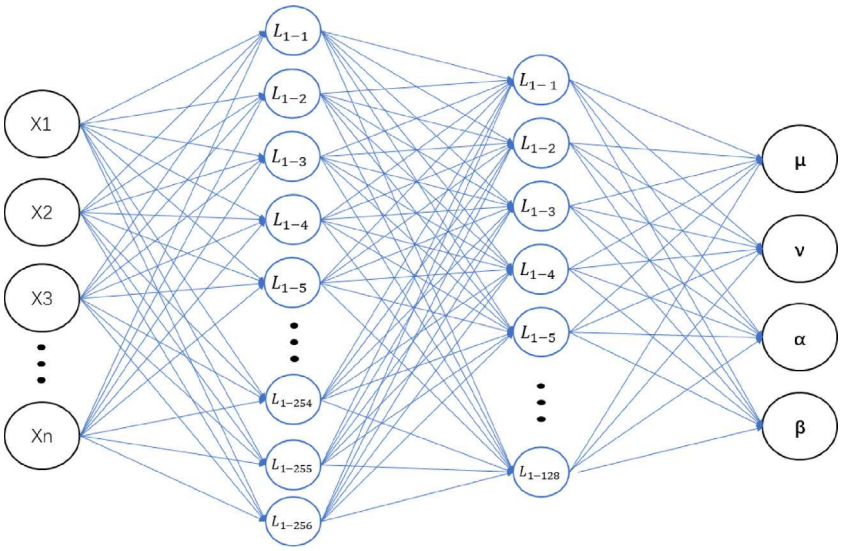

Fig. 1. Neural Network Structure.

2.2. Prediction framework definition

In this paper, we use a data-driven framework based on the deep evidential model that combines the empirical models with high-resolution, accelerometer-inferred densities from the CHAMP satellite, and other geomagnetic and solar indices. The inputs and the output of the model can be described as Eq. (5), which has been studied in our previous papers [25,26]

$$\lg\left(\hat{\rho}\left(t\right)\right) = f\begin{pmatrix} \lg\left(\rho_{JB}(t)\right), \ldots \lg\left(\rho_{JB}\left(t - D_{JB}t_s\right)\right) \\ \lg\left(\rho_{NRL}(t)\right), \ldots \lg\left(\rho_{NRL}\left(t - D_{NRL}t_s\right)\right) \\ F_{10.7}(t-1d), F_{10.7A}(t-1d) \\ Ap(t), F_{30}(t), Dst(t), SymH(t) \\ \rho_{CHAMP}(t - t_D) \end{pmatrix} \tag{5}$$

The $\hat{\rho}$ on the left side of the equation is the predicted density from the evidential model at 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 h. $Ap(t)$ 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. $SymH(t)$ is the one-minute resolution version of the Dst index [30]. $\rho_{CHAMP}(t)$ is the density value referred from CHAMP 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. There is also a time delay t_D in the CHAMP density measurement, which is set as 300 s. Our previous studies have demonstrated that it is necessary to provide these inputs in order 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 [31]. The estimated density derived from the NRLMSISE-00 model was obtained from [32]. 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 [33]. To access the Dst data, we refer to [34], while the $Symh$ data can be obtained from [35]. For the density of the CHAMP satellite, we refer to Mehta et al. [36,37]. Here we make the assumption that the density from the CHAMP 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 [18–20,24].

2.3. Neural network structure

The neural network is built based on the toolbox Keras [38] in Python 3.9. We first optimize the neural network structures using the KerasTuner [39] and then make further modifications to improve the results. The neural network structure can be visualized in Fig. 1.

The architecture contains 256 neurons in the first hidden layer, 128 neurons in the second hidden layer, and four outputs, which are the hyperparameters for the evidential distribution. The activation function used in the network is linear, with a batch size of 256, and 500 epochs for training. We normalize the data before training, and the same neural network structure is used for both the quiet and storm conditions.

2.4. 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):

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

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

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 is a measure of 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 range of confidence interval is defined as $CL = [5\%, 10\%, \ldots, 95\%, 99\%]$. The corresponding coefficients defining the uncertainty bounds are then given as Eq. (9).

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

where erf is the error function, defined as Eq. (10):

$$\mathrm{erf}\, x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{10}$$

To evaluate the reliability of the uncertainty, we calculate the mean absolute calibration error (MACE), which is defined as Eq. (11).

$$MACE = \frac{1}{n_C} \sum_{k=1}^{n_C} |C[k] - P[k]| \tag{11}$$

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 [40] to the standard deviation of the predicted value to help improve the overall calibration of the model. The scalar factor is calculated by Eq. (12).

$$s = \sqrt{\frac{1}{n_v} \sum_{i=1}^{n_v} \left[\frac{\left\| \mathbf{y_i} - \hat{\mu}_\mathbf{i} \right\|^2}{\hat{\sigma}_i^2} \right]} \tag{12}$$

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 ith validation sample.

3. Case studies and results

3.1. Quiet time

The data used in this quiet time study is selected from the year 2007. The training database is chosen from 04/14/2007 to 06/30/2007. There are two reasons for the selection of this period. First, the smallest Dst during the period is −63 nT, which is larger than the definition of the intense storm which we will study later. By focusing on a period that can be considered as "quiet time", we ensure that other factors do not confound the results. Second, this is the period that has been studied in [24,25], which uses GPs as the underlying machine learning model. Therefore, the performance can now be compared between the evidential model and the GPs model.

In order to evaluate the model's performance when the test data is both within and outside of the training range, two distinct test sections are designed. The first test section includes data from the same time period as the training data, providing insight into the model's performance when applied to the training data range. The second test section is selected from 07/01/2007 to 07/31/2007, which is not included in the training data and represents a future time period. The purpose of this design is to show that the model is able to generalize and can make predictions outside the training data.

We present the numerical results as Table 1 and compare them with the results from the GPs model [26]. The evidential model shows higher accuracy and a more reliable uncertainty prediction than the GPs model, as the results from the evidential model show larger R and coverage rate values and smaller RMSE and MACE values than the GPs model. The results also indicate that when the test section is within the training data range, the accuracy of the predicted results is better than that of the test data that is out of the training range, which is what we expect.

Table 1
Quiet time: Numerical result.

Model	Evi-OutRange	GPs-OutRange [26]	Evi-InRange	GPs-InRange
R	0.9027	0.9019	0.9495	0.9490
RMSE $\times 10^{-12}$	0.2616	0.2688	0.2958	0.3006
Coverage Rate	0.9388	0.9160	0.9851	0.9849
MACE	0.0062	0.0087	0.0091	0.0148

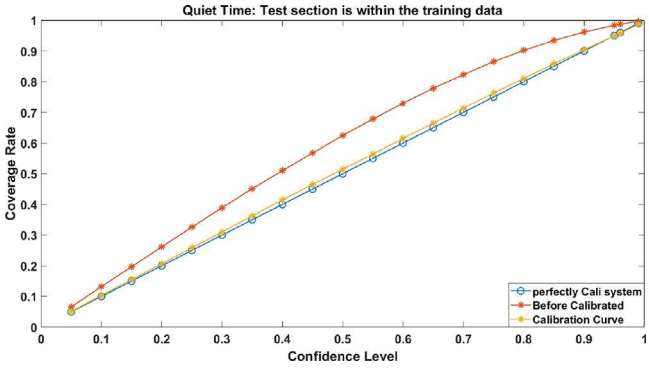

Fig. 2. Quiet Time - Recalibration Curve for test section within the training range. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

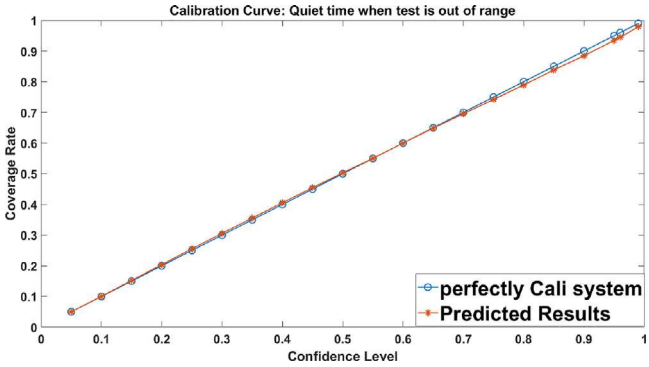

Fig. 3. Quiet Time - calibration Curve for test section out of training range.

We plot the calibration curve with and without the re-scale factor in Fig. 2. The blue line is the perfectly calibrated curve, the red line is the predicted calibration curve without the re-scale factor, and the yellow line is the calibration curve with the re-scale factor. We can clearly see that the uncertainty prediction has been improved with the re-scale factor.

Based on the re-calibrated results, we plot the calibration curve for the test section that is out of the range in Fig. 3. The calibration curve shows a minimal gap compared to the perfectly calibrated line, and the MACE value is also very close to zero. These indicate that the uncertainty provided by the evidential model for the quiet time test case is reliable and trustworthy.

We plot the uncertainty boundaries with the corresponding predicted results for the two conditions in Fig. 4. The first subplot shows the predicted results when the test section is within the training range, while the second subplot shows the results when the test section is outside the training range. The red section represents the actual values, the blue section represents the predicted values, and the lighter blue area indicates the uncertainty boundary.

By comparing the two subplots in Fig. 4, we can see that when the test data is out of the training range, the uncertainty range is more expansive, indicating larger uncertainty in the predictions. The second subplot shows some extreme values, which are not present in the first

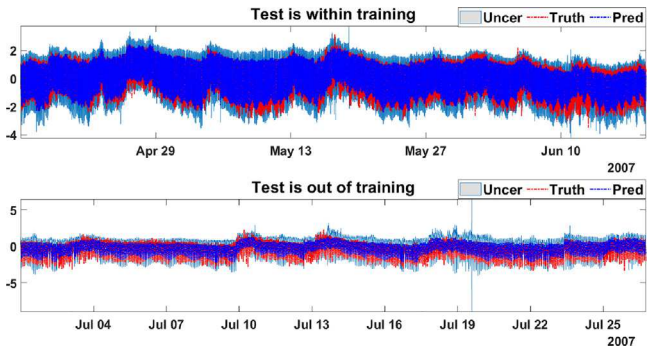

Fig. 4. Quiet Time - Predicted results for the whole period with corresponding 2σ uncertainty boundary. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

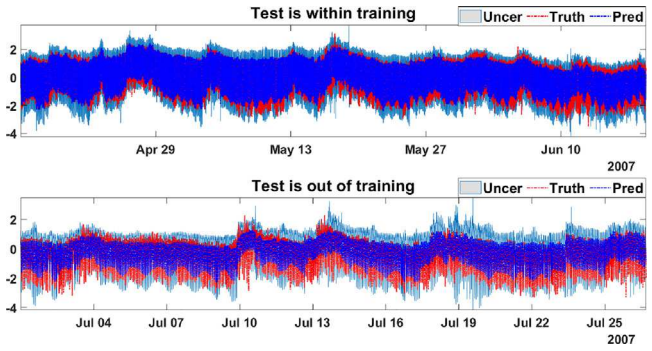

Fig. 5. Quiet Time - zoomed in predictions.

subplot, highlighting the effect of the test data when it is not included in the training range.

To get a better idea of the overall trend, we zoom in on the y-axis to exclude those extra large values, as Fig. 5 shows. It is now convenient to compare the two test cases as the y-value range is the same for both plots. Since both parts correspond to a quiet period of 2007, the true density value (in red) does not significantly change over time. The range of predicted values and the associated uncertainties are similar in both cases. However, by observing the predicted values, we can see that the predicted results are closer to the true values in the first subplot. This is because the test data in the first subplot can be considered as a validation part, while the second test data is a new, unused dataset that is out of the training range.

We select a more specific short period on the two conditions for a more detailed analysis, and the results are as Fig. 6 presents. When the test is out of training range, the uncertainty boundaries on the second subplot show very large fluctuations around July 13th, 16:00.

We plot the aleatoric and epistemic uncertainties along the Dst for the two conditions. In Fig. 7 we show the uncertainties when the test section is within the training range, and in Fig. 8 we show the uncertainties when the test section is out of range. The red line represents the aleatoric uncertainty, and the yellow line is the epistemic uncertainty. From Figs. 7 and 8 we can see the uncertainties from the second condition have a wider range. Extreme values appear in the test section that is out of the training range, which makes it hard to see the uncertainty distribution clearly. To reduce this effect, we zoom in on the y-axis while excluding the extreme values, resulting in the plotted results shown in Fig. 9.

Comparing Fig. 7 with Fig. 9, we can see the ranges of the aleatoric uncertainty value for the two test cases are very close, except for the extreme values that have been excluded. The mean values of the aleatoric uncertainty for both cases are around 0.3.

Fig. 6. Quiet Time - Predictions in a short period.

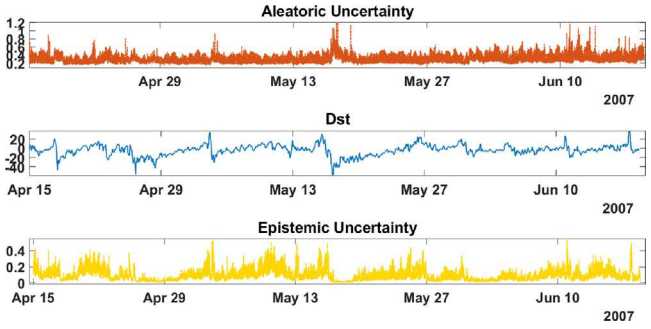

Fig. 7. Quiet Time - The uncertainty distribution when the test is within training range. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

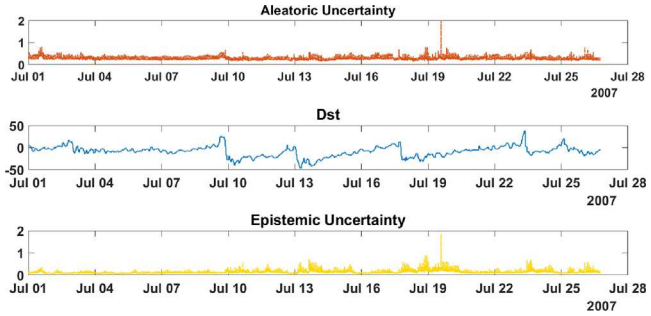

Fig. 8. Quiet Time - The uncertainty distribution when the test is out of training range. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

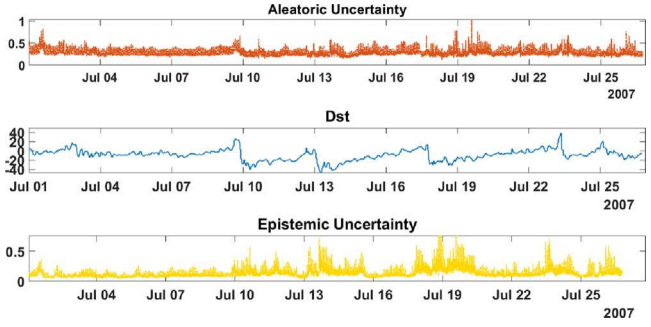

Fig. 9. Quiet Time - zoomed in uncertainty distribution when the test is out of training range.

Table 2
Storm period.

Year	Date	Min Dst	Year	Date	Min Dst
2003	5.29–6.2	−164	2005	1.7-1.8	−112
	6.18–6.19	−165		1.18-1.19	−107
	8.18–8.19	−140		5.8-5.9	−117
	10.29-10.30	−432		5.15-5.16	−305
	11.20–11.21	−490		5.30-5.31	−127
2004	1.22–1.23	−137		6.13-6.14	−113
	2.11–2.12	−107		6.23-6.24	−101
	3.10–3.11	−101		7.10-7.11	−114
	4.3–4.4	−149		**8.24-8.25**	−179
	7.22–7.28	−208		**8.31-9.1**	−119
	8.30–8.31	−128		**9.11-9.12**	−147
	11.8–11.11	−397			

Table 3
Test cases list.

Index	Period
Case-A	10/29/2003-10/31/2003
Case-B	08/24/2005-08/25/2005
Case-C	08/31/2005-09/01/2005
Case-D	09/11/2005-09/12/2005

In Fig. 7, the Dst for the first condition displays a declining pattern on April 26th, May 4th, May 17th, and June 12th. In Fig. 9, a rapid reduction in Dst was observed on July 10th, July 13th, July 19th, and July 23rd. The changes in Dst correspond to fluctuations of both uncertainties. During significant Dst changes, both the aleatoric and epistemic uncertainties rise for a short period. During the quiet time, both uncertainties exhibit minimal variations, though the aleatoric uncertainty shows more stability with smaller fluctuations.

During the quiet time, the aleatoric uncertainty generated by the evidential model exhibits similar performances for both conditions. The mean values of the aleatoric uncertainty are very similar regardless of whether the test data falls within or outside the training range. The two test cases also show a noticeable difference in epistemic uncertainty. In cases with out-of-range test data, the mean value of the epistemic uncertainty is larger than the value when the test data is within the range. In general, we can conclude that the evidential model can provide reliable uncertainty estimations and high-accuracy predictions during quiet time, even when the test data is out of the training data range.

3.2. Storm time

To study the model's performance for storm situations, we find the periods when the Dst is smaller than −100 nT from the years 2003 to 2005. The storm conditions are summarized in Table 2 with the minimum Dst during the corresponding periods. The training data is selected from the years 2003, 2004, and the first half of 2005 until the end of July. The bold dates (except 10/29/2003 - 10/30/2003) are used as the test section not included in the training database, which are summarized in Table 3.

Case-A is when the test section is from 10/29/2003 to 10/30/2003, which covers the 2003 Halloween storm and contains the second biggest storm in 2003.

Case-B is when the test section is from 08/24/2005 to 08/25/2005, which contains the second biggest storm in 2005. The training data is defined as Table 2, from the beginning of 2003, until 07/11/2005. This test case is not included in the training section but is in the future time of the training data.

Case-C is when the test section is from 08/31/2005 to 09/01/2005. The Dst distribution, in this case, is similar to Case-B, but the minimum value is larger than Case-B. The test section is not included in the training data, but the test section is in a further future time than the training section.

In Case-D, the test section is during 09/11/2005 and 09/12/2005. The test section is two months later than the last training date and is the furthest future date from the training date.

Table 4
Case-A: Numerical results.

Model	Case-A1	Case-A2	GPs Case-A1 [26]	GPs Case-A2 [26]
R	0.8175	0.8086	0.8123	0.7986
RMSE $\times 10^{-12}$	2.0629	2.1079	2.1130	2.1181
Coverage Rate	0.9851	0.9907	0.9199	0.8950
MACE	0.0061	0.0072	0.1112	0.1154

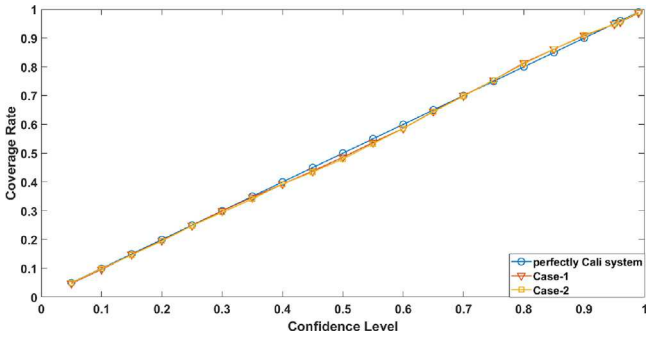

Fig. 10. Case-A: Calibration Curve. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2.1. Case-A

In this case, we will show two conditions: One is when the test section is included in the training data, which we call "Case-A1" for short. The other case is when the test section is not included in the training sections, which means the training data are the selected storms from the beginning of 2003 to the first half of 2005, except the period from 10/29/2003 to 10/30/2003. To show the difference, we call this "Case-A2".

The numerical results of the two conditions are presented in Table 4, and we compare the results from the evidential model with results from the (Gaussian Processes) GPs model [26].

Table 4 indicates that when the storm period is excluded from the training data, the prediction accuracy in Case-A2 is inferior to that in Case-A1. Despite this, the result obtained from the evidential model still achieves better results compared to the results obtained from the GPs model used in [26]. In the evidential model, the coverage rate increases significantly compared with the GPs model. The MACE value from the evidential model also shows advantages over the GPs model.

We plot the calibration curve for the Case-A1 and Case-A2 of the evidential model in Fig. 10. The difference in the calibration curves between Case-A1 (Red) and Case-A2 (yellow) is small.

We plot the aleatoric and epistemic uncertainties from Case-A1 in the first subplot and Case-A2 in the third subplot with the Dst in the second subplot in Fig. 11, with the blue line representing total uncertainty, the red line representing the aleatoric uncertainty, and the yellow line representing the epistemic uncertainty.

The distributions of the uncertainties in the two cases are very similar. There are two distinct minima of the Dst according to Fig. 11, revealing there are two storms happened during the period from 10/29/2003 to 10/30/2003. The first storm occurred from Oct. 29th, 14:00 to Oct.30, 17:30, then followed by the second storm which ended around Oct. 31st, 11:00. In Case-A1, when the test data is included in the training range, the mean value of the aleatoric uncertainty during the first storm is 0.4694, and the mean aleatoric value in the second storm is 0.4028. In Case-A2, when the test data is not in the training

Table 5
Case-B: Numerical results.

Model	Evidential	GPs [26]
R	0.8992	0.8985
RMSE $\times 10^{-12}$	1.3901	1.3926
Coverage Rate	0.9739	0.9057
MACE	0.0113	0.0639

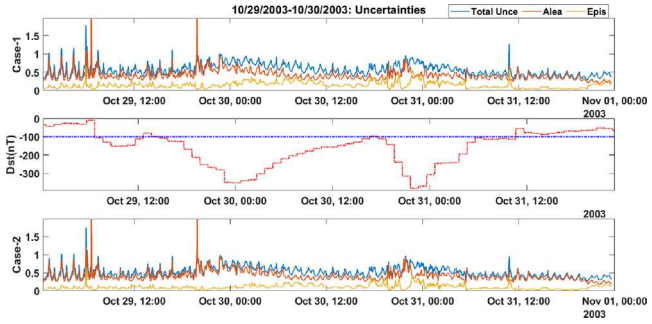

Fig. 11. Case-A: Uncertainty Distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

range, the aleatoric mean during the first storm is 0.4711, and the second storm is 0.4114. For the epistemic uncertainty, on the other hand, the mean value during the first storm in Case-A1 is 0.1089, and 0.1319 for the second storm. While in Case-A2, the mean value of the epistemic uncertainty during the first storm is 0.1661, and for the second storm the value is 0.1762. Except for the storm period, the mean of aleatoric during the quiet time from both Case-A1 and Case-A2 is close to 0.3, which is close to the mean aleatoric value during the quiet time studied in the last subsection.

In this test case, Case-A1 performs better than Case-A2, showing larger R and coverage rate values and smaller RMSE and MACE values, although the differences are small. Analyzing the aleatoric and epistemic uncertainty distributions during the storm period, we can see that the main uncertainty is from the aleatoric uncertainty and the total uncertainty increases during the storm period. The possible reason why the aleatoric uncertainty became more extensive during the storm period is that the suddenly happened magnetic storm is a more unpredictable and chaotic process which is what the aleatoric uncertainty is designed to capture.

3.2.2. Case-B

The numerical results for this case are presented in Table 5 and compared with the results from the GPs model.

The evidential model gives a better prediction on the accuracy, as its R is larger than the GPs model, and the RMSE is smaller. Additionally, the evidential model exhibits a higher coverage rate and a smaller MACE than the GPs model. We plot the calibration curve for the evidential model in Fig. 12.

The calibrated curve of the predicted results is very close to the perfectly calibrated curve. The MACE from the evidential model is smaller than the value from the GPs model, which indicates that the evidential model can provide a more reliable uncertainty than our previous model.

We plot the predicted value and the uncertainty boundaries with the corresponding Dst along the time axis, as is shown in Fig. 13.

From Fig. 13, we can see most of the truth can be covered within the uncertainty boundary. It is also interesting to see that during the storm period, the value of the uncertainty increase. Around Aug. 24th, 11:00, an extremely large uncertainty occurs, which corresponds to the time the smallest Dst happened.

We present the distribution of the aleatoric and epistemic uncertainties, as shown in Fig. 14 along the Dst.

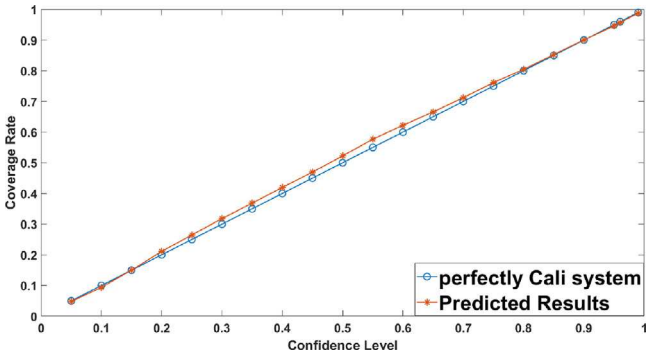

Fig. 12. Case-B: Calibration Curve.

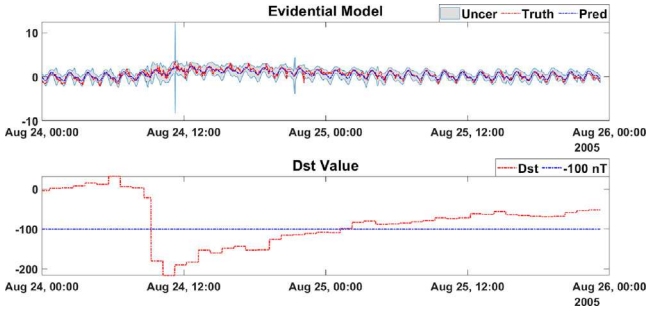

Fig. 13. Case-B: Predicted results.

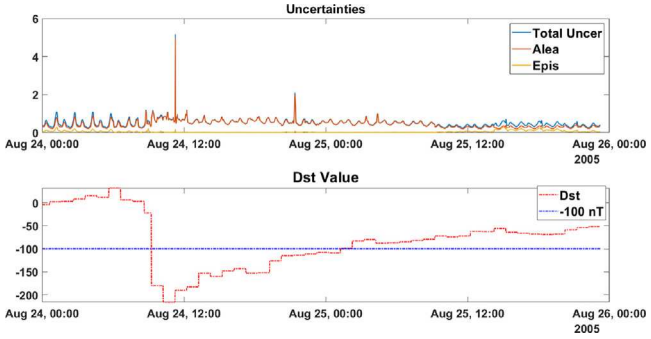

Fig. 14. Case-B - Uncertainty Distribution.

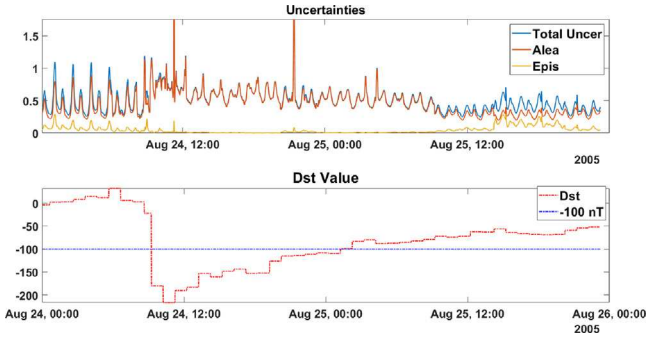

Fig. 15. Case-B - Zoomed in Uncertainty Distribution.

From Fig. 14 we can see the distribution of the aleatoric uncertainty shows some extreme values during the storm period. To see the two uncertainties more clearly, we zoom in from the y-axis to ignore the outliers of the large values, and a more detailed plot is shown in Fig. 15.

Table 6
Case-C: Numerical results.

Model	Evidential	GPs [26]
R	0.8874	0.8840
RMSE $\times 10^{-12}$	0.7026	0.7020
Coverage Rate	0.9961	0.9659
MACE	0.0128	0.0833

Table 7
Case-D: Numerical results.

Model	Evidential	GPs [26]
R	0.8067	0.7803
RMSE $\times 10^{-12}$	1.0915	1.1946
Coverage Rate	0.9859	0.9465
MACE	0.0355	0.0837

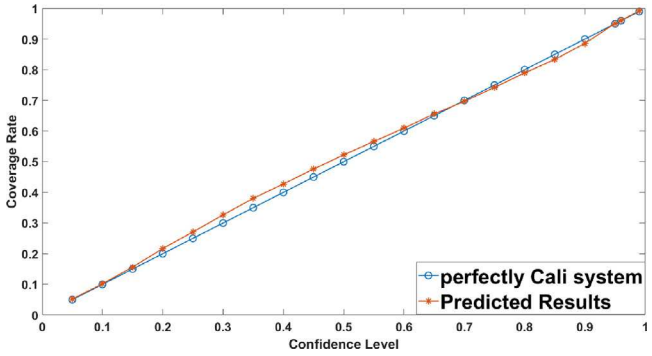

Fig. 16. Case-C: Calibration Curve. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

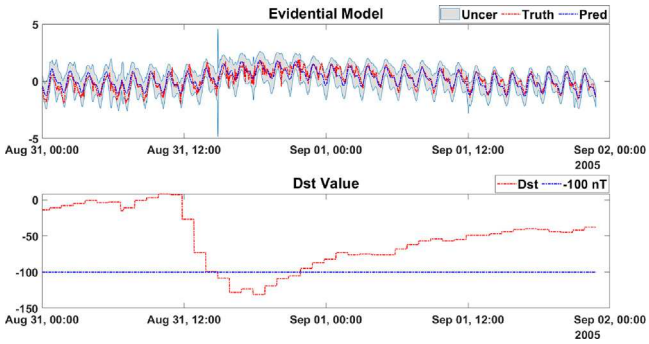

Fig. 17. Case-C: Predicted results.

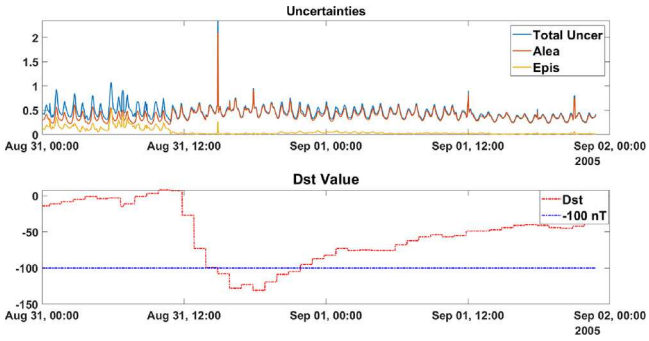

Fig. 18. Case-C - Uncertainty Distribution.

As presented in Fig. 15, the total uncertainty becomes larger during the storm time. On the one hand, the aleatoric uncertainty value remains relatively stable when the Dst index is larger than −100 nT, with a mean value of approximately 0.3, which is close to the mean aleatoric value observed during quiet periods. However, when the Dst index drops below −100 nT, around 09:00 on Aug. 24th, the aleatoric uncertainty value increases significantly, remaining elevated until the storm event ends at 02:00 on Aug. 25th. On the other hand, the epistemic uncertainty remains relatively low until 14:00 on Aug. 25th, when the Dst goes up to −50 nT, which is commonly referred to as a quiet time. Overall, we can see that the evidential model effectively distinguishes between the data obtained during the storm and quiet periods. And the model correctly outputs that its prediction uncertainties are larger during the storm period.

3.2.3. Case-C

The predicted results in Case-C are presented in Table 6 and compared with the GPs model.

The numerical results obtained from this period demonstrate that the evidential model provides an improvement in performance compared to the GPs model. The evidential model is able to produce more accurate predictions and a better representation of the uncertainty associated with these predictions.

We plot the calibration curve of this test case in Fig. 16. The blue line is the perfectly calibrated system, and the red line is the coverage rate from the predicted results.

As shown in the curve, the predicted coverage rate initially exceeds the perfect line but then drops below it when the confidence level exceeds 0.7. However, when the coverage rate reached above 95%, the predicted coverage rate converged to the perfect line. The MACE value is also very close to zero, which indicates the estimations from the evidential model are reliable.

Fig. 17 shows the truth value, predicted results, and uncertainty boundaries along with the Dst in time series.

The evidential model can capture the distribution of the true density well because the predicted results are very close to the true data. The uncertainty boundaries can cover most of the truth data, leading to a high coverage rate. The uncertainty shows a very large value when the time is around Aug. 31st, 12:48, which is close to the time the storm begins. We show the distributions of the aleatoric and epistemic uncertainties in this case in Fig. 18.

From Fig. 18 we can see that the storm begins around Aug. 31st, 14:00, with an increment of the aleatoric uncertainty and a decrease in epistemic uncertainty. Before Aug. 31st, 14:00, and after Sep. 1st, 00:00, the mean value of the aleatoric uncertainty during the quiet period in this storm case is 0.3, the same as the mean aleatoric uncertainty during the quiet time test case. However, unlike the aleatoric uncertainty changing along the intense storm change, the epistemic uncertainty remains relatively low until the Dst goes up −50 nT.

3.2.4. Case-D

The predicted results are listed in Table 7. The evidential model gives more accurate predictions and more reliable uncertainty estimations compared to the GPs model.

The calibration curve of the storm case is plotted in Fig. 19. We can see the predicted results (red) shows a larger coverage rate than the perfectly calibrated system line (blue). The coverage rate from the predicted value is firstly smaller than the ideal value and then becomes larger.

The predicted results and the corresponding uncertainty boundaries are plotted in Fig. 20, with the corresponding Dst during the same period. For this case, we can see the storm period is between Sep. 10th, 05:00 to Sep. 11th, 14:00. The total uncertainty shows an extremely large value during this period.

We also plot the uncertainties in this storm case in Fig. 21. When we look at the two uncertainties during this test case, we can see that the epistemic uncertainty becomes smaller during the storm period.

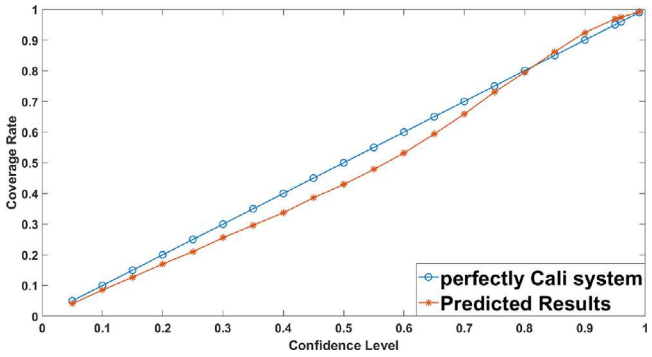

Fig. 19. Case-D: Calibration Curve. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

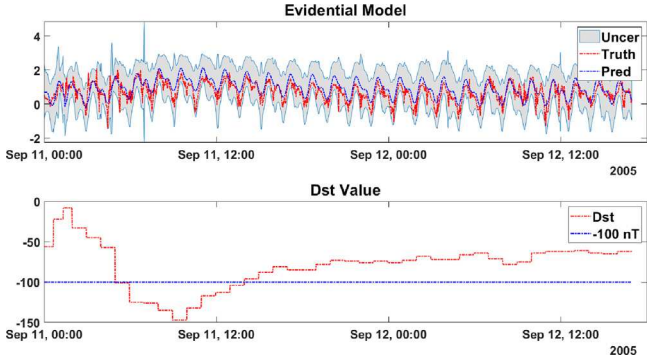

Fig. 20. Case-D: Predicted results.

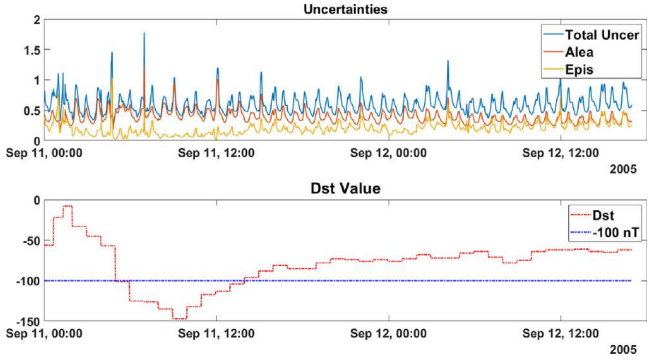

Fig. 21. Case-D - Uncertainty Distribution.

Meanwhile, the aleatoric uncertainty becomes slighter larger than the other time during the storm period. When the Dst are larger than −100 nT, the aleatoric values are around 0.3, which is very close to the aleatoric uncertainty during the quiet time.

3.3. Aleatoric and epistemic uncertainties for all the cases

We calculate the mean value of the aleatoric and epistemic uncertainties for all the cases that have been studied. "Quiet Time" in the second row refers to the test case in 2007 when the minimum Dst during this period was larger than −100 nT. Then we present the uncertainty mean values in the storm cases. For Case-A1 and A2, the 1st storm refers to the period from Oct. 29th, 14:00 to Oct.30, 17:30, which is followed by the 2nd storm that ends around Oct. 31st, 11:00. For all the storm cases, "storm period" refers to the duration when the Dst is smaller than −100 nT and "quiet period" in storm cases refers to the duration before the storm begins and after it subsides (see Table 8).

Table 8
Mean values of aleatoric and epistemic uncertainties.

	Alea	Epis
Quiet Time	0.3064	0.0844
Case-A1 - 1st Storm	0.4694	0.1089
Case-A1 - 2nd Storm	0.4028	0.1319
Case-A1 - quiet period	0.3106	0.1003
Case-A2 - 1st Storm	0.4711	0.1611
Case-A2 - 2nd Storm	0.4114	0.1762
Case-A2 - quiet period	0.3179	0.1104
Case-B - storm period	0.5792	0.0218
Case-B - quiet period	0.3155	0.1207
Case-C - storm period	0.4582	0.0296
Case-C - quiet period	0.3029	0.1678
Case-D - storm period	0.4466	0.1162
Case-D - quiet period	0.3175	0.1189

Table 8 reveals that the mean value of the aleatoric uncertainty during quiet time is around a specific value (0.3) and the aleatoric mean during the storm period is larger than during the quiet period. As for the epistemic uncertainty, the mean value is around 0.1 during the quiet period. However, the epistemic value does not show a clear pattern during the storm periods. The epistemic mean value is smaller during the storm period for the last three storm cases. While for the first storm case, because the double storm condition is more complicated and unpredictable, the epistemic value shows fluctuations during the storm and the mean values are larger than during the quiet period.

4. Conclusion

In this paper, we propose using an evidential model-based framework to predict the density of the atmosphere during periods of both the quiet time in 2007 and the storm cases in 2003 and 2005. Study results show higher R values and smaller RMSE values than our previous framework based on the Gaussian Processes (GPs). For the quiet time, when the test case is within the training data range, the R value reaches 0.9495. When the test data is out of the training range, which is in the future training dates, the R value can reach 0.9027. For the storm cases we studied in the paper, the R value is always larger than 0.80, which is better than our previous study based on the GPs model. The best R value of the storm from 08/31/2005 to 09/01/2005 reaches 0.8874, with the corresponding RMSE value as 0.7026×10^{-12} while the test period is one and a half months later than the training database.

Additionally, the uncertainty boundaries from the proposed framework can cover most of the actual data. The coverage rate is always above 93% and with a small Mean Absolute Calibration Error (MACE) value, indicating a high degree of confidence in the model's predictions. This suggests that the proposed model provides accurate predictions and a robust representation of the uncertainty associated with the predictions.

In summary, the proposed deep evidential model-based framework is shown to provide accurate predictions with reliable uncertainties for both quiet and storm periods. The evidential model also provides insight into both aleatoric and epistemic uncertainties, which can be used to analyze the data and the reliability of the models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research has been supported by National Science Foundation, United States, under award number 2149747.

References

[1] F.A. Marcos, M. Rendra, J. Griffin, J. Bass, D. Larson, J. Liu, Precision low earth orbit determination using atmospheric density calibration, J. Astronaut. Sci. 46 (1998) 395–409.

[2] D.M. Oliveira, E. Zesta, H. Hayakawa, A. Bhaskar, Estimating satellite orbital drag during historical magnetic superstorms, Space Weather 18 (11) (2020) e2020SW002472.

[3] J. Picone, A. Hedin, D.P. Drob, A. Aikin, NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues, J. Geophys. Res. Space Phys. 107 (A12) (2002) SIA–15.

[4] B. Bowman, W.K. Tobiska, F. Marcos, C. Huang, C. Lin, W. Burke, A new empirical thermospheric density model JB2008 using new solar and geomagnetic indices, in: AIAA/AAS Astrodynamics Specialist Conference and Exhibit, 2008, p. 6438.

[5] NASA, Earth atmosphere model - imperial units, NASA, https://www.grc.nasa.gov/www/k-12/airplane/atmos.html.

[6] C.W. Group, Cospar International Reference Atmosphere–2012, Tech. Rep. Technical Report, The Committee on Space Research, 2012.

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

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

[9] A. Burns, T. Killeen, W. Deng, G. Carignan, R. Roble, Geomagnetic storm effects in the low to middle latitude upper thermosphere, J. Geophys. Res. Space Phys. 100 (A8) (1995) 14673–14691.

[10] A. Burns, T. Killeen, W. Wang, R. Roble, The solar-cycle-dependent response of the thermosphere to geomagnetic storms, J. Atmos. Sol.-Terr. Phys. 66 (1) (2004) 1–14.

[11] W.J. Burke, C.Y. Huang, F.A. Marcos, J.O. Wise, Interplanetary control of thermospheric densities during large magnetic storms, J. Atmos. Sol.-Terr. Phys. 69 (3) (2007) 279–287.

[12] D.M. Oliveira, E. Zesta, Satellite orbital drag during magnetic storms, Space Weather 17 (11) (2019) 1510–1533.

[13] H. Demars, R. Schunk, Seasonal and solar cycle variations of the polar wind, J. Geophys. Res. Space Phys. 106 (A5) (2001) 8157–8168.

[14] W. Gonzalez, J.-A. Joselyn, Y. Kamide, H.W. Kroehl, G. Rostoker, B. Tsurutani, V. Vasyliunas, What is a geomagnetic storm? J. Geophys. Res. Space Phys. 99 (A4) (1994) 5771–5792.

[15] W.D. Gonzalez, B.T. Tsurutani, A.L.C. De Gonzalez, Interplanetary origin of geomagnetic storms, Space Sci. Rev. 88 (3) (1999) 529–562.

[16] Y. Zhou, S. Ma, H. Lühr, C. Xiong, C. Reigber, An empirical relation to correct storm-time thermospheric mass density modeled by NRLMSISE-00 with CHAMP satellite air drag data, Adv. Space Res. 43 (5) (2009) 819–828.

[17] R. Liu, H. Lühr, E. Doornbos, S.-Y. Ma, Thermospheric mass density variations during geomagnetic storms and a prediction model based on the merging electric field, Ann. Geophys. 28 (9) (2010) 1633–1645.

[18] D. Pérez, B. Wohlberg, T.A. Lovell, M. Shoemaker, R. Bevilacqua, Orbit-centered atmospheric density prediction using artificial neural networks, Acta Astronaut. 98 (2014) 9–23.

[19] C. Xiong, H. Lühr, M. Schmidt, M. Bloßfeld, S. Rudenko, An empirical model of the thermospheric mass density derived from CHAMP satellite, Ann. Geophys. 36 (4) (2018) 1141–1152.

[20] S. Bonasera, G. Acciarini, J.A. Pérez-Hernández, B. Benson, E. Brown, E. Sutton, M.K. Jah, C. Bridges, A.G. Baydin, Dropout and ensemble networks for thermospheric density uncertainty estimation.

[21] D.J. Gondelach, R. Linares, P.M. Siew, Atmospheric density uncertainty quantification for satellite conjunction assessment, J. Guid. Control Dyn. 45 (9) (2022) 1760–1768.

[22] R.J. Licata, P.M. Mehta, W.K. Tobiska, S. Huzurbazar, Machine-learned HASDM thermospheric mass density model with uncertainty quantification, Space Weather 20 (4) (2022) e2021SW002915.

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

[24] T. Gao, H. Peng, X. Bai, Calibration of atmospheric density model based on Gaussian processes, Acta Astronaut. 168 (2020) 273–281.

[25] Y. Wang, H. Peng, X. Bai, J.T. Wang, H. Wang, Advance thermospheric density predictions through forecasting geomagnetic and solar indices based on Gaussian processes, in: AAS/AIAA Astrodynamics Specialist Conference, 2021.

[26] Y. Wang, X. Bai, Comparison of Gaussian processes and neural networks for thermospheric density predictions during quiet time and geomagnetic storms, in: AAS/AIAA Astrodynamics Specialist Conference, 2022.

[27] S.C. Hora, Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management, Reliab. Eng. Syst. Saf. 54 (2–3) (1996) 217–223.

[28] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2) (2009) 105–112.

[29] A. Amini, W. Schwarting, A. Soleimany, D. Rus, Deep evidential regression, Adv. Neural Inf. Process. Syst. 33 (2020) 14927–14937.

[30] J.A. Wanliss, K.M. Showalter, High-resolution global storm index: Dst versus SYM-H, J. Geophys. Res. Space Phys. 111 (A2) (2006).

[31] M. Mahooti, Jacchia-Bowman atmospheric density model, The MathWorks Inc., https://www.mathworks.com/matlabcentral/fileexchange/56163-jacchia-bowman-atmospheric-density-model.

[32] M. Mahooti, NRLMSISE-00 atmosphere model, The MathWorks Inc., https://www.mathworks.com/matlabcentral/fileexchange/56253-nrlmsise-00-atmosphere-model.

[33] T.S. Kelso, Space weather data documentation, CelesTrak, https://celestrak.org/SpaceData/SpaceWx-format.php.

[34] K.-O. Cho, Geomagnetic equatorial DST index home page, World Data Center for Geomagnetism, Kyoto, https://wdc.kugi.kyoto-u.ac.jp/dstdir/.

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

[36] P.M. Mehta, A.C. Walker, E.K. Sutton, H.C. Godinez, New density estimates derived using accelerometers on board the CHAMP and GRACE satellites, Space Weather 15 (4) (2017) 558–576.

[37] CHAMP and GRACE density data sets, https://drive.google.com/drive/folders/0BwtX8XEH-aEueHJiU1htLVo0cms?resourcekey=0-byxPMLbZSVC5Blxb_Rfl9g.

[38] F. Chollet, et al., Keras, 2015, https://keras.io.

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

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