Space Weather®

RESEARCH ARTICLE
10.1029/2021SW002915

Machine-Learned HASDM Thermospheric Mass Density Model With Uncertainty Quantification

Richard J. Licata[1], Piyush M. Mehta[1], W. Kent Tobiska[2], and S. Huzurbazar[3]

[1]Department of Mechanical and Aerospace Engineering, West Virginia University, Morgantown, WV, USA, [2]Space Environment Technologies, Pacific Palisades, CA, USA, [3]School of Mathematical and Data Sciences, West Virginia University, Morgantown, WV, USA

Key Points:

- Thermosphere model based on High Accuracy Satellite Drag Model (HASDM) outputs is developed with robust and reliable uncertainty estimates
- HASDM-ML provides predictions with approximately 10% error relative to HASDM database across 20 years of available historical data
- Gaussian loss function provides model with well-calibrated uncertainty estimates across diverse space weather conditions

Correspondence to:

R. J. Licata,
rjlicata@mix.wvu.edu

Citation:

Licata, R. J., Mehta, P. M., Tobiska, W. K., & Huzurbazar, S. (2022). Machine-learned HASDM thermospheric mass density model with uncertainty quantification. *Space Weather*, *20*, e2021SW002915. https://doi.org/10.1029/2021SW002915

Received 14 SEP 2021
Accepted 5 MAR 2022

Abstract A thermospheric neutral mass density model with robust and reliable uncertainty estimates is developed based on the Space Environment Technologies (SET) High Accuracy Satellite Drag Model (HASDM) density database. This database, created by SET, contains 20 years of outputs from the U.S. Space Force's HASDM, which currently represents the state of the art for density and drag modeling. We utilize principal component analysis for dimensionality reduction, which creates the coefficients upon which nonlinear machine-learned (ML) regression models are trained. These models use three unique loss functions: Mean square error (MSE), negative logarithm of predictive density (NLPD), and continuous ranked probability score. Three input sets are also tested, showing improved performance when introducing time histories for geomagnetic indices. These models leverage Monte Carlo dropout to provide uncertainty estimates, and the use of the NLPD loss function results in well-calibrated uncertainty estimates while only increasing error by 0.25% (<10% mean absolute error) relative to MSE. By comparing the best HASDM-ML model to the HASDM database along satellite orbits, we found that the model provides robust and reliable density uncertainties over diverse space weather conditions. A storm-time comparison shows that HASDM-ML also supplies meaningful uncertainty estimates during extreme geomagnetic events.

Plain Language Summary An upper-atmospheric density model with robust and reliable uncertainty estimates is developed based on the Space Environment Technologies High Accuracy Satellite Drag Model (HASDM) density database. This database contains 20 years of outputs from HASDM, which represents the state of the art for density and drag modeling. We use a decomposition tool called principal component analysis to reduce the dimensionality of the data set. Three loss functions, mean square error, negative logarithm of predictive density (NLPD), and continuous ranked probability score, are tested with three input sets to identify the best-performing model. We optimize nine models (all three loss functions and input sets) and compare the prediction accuracy and the reliability of its uncertainty estimates. The models leverage Monte Carlo dropout to generate probabilistic outputs from which we extract model uncertainty. We find that using an input set containing a time series for the geomagnetic indices results in the most accurate models. In addition, the model using these inputs with the NLPD loss function has sufficient performance (approximately 10% absolute error) and the most calibrated/reliable uncertainty estimates on independent data. We test this model's uncertainty capabilities in the density space along satellite orbits from 2002 to 2010 showing the model's reliability across diverse space weather conditions.

1. Introduction

Space Situational Awareness is a major focus of space agencies and private defense/technology companies worldwide. With the number of objects in low-Earth orbit (LEO) continuously growing, knowledge of future satellite/debris positions is becoming increasingly important (Radtke et al., 2017). While there are numerous perturbations affecting the trajectories of objects, atmospheric drag is the largest source of uncertainty in the LEO region (Emmert et al., 2017; Storz et al., 2005). Our current understanding of the thermosphere is incomplete, resulting in imperfect modeling of neutral mass density. Over the past several decades, researchers have developed increasingly accurate models and made improvements to existing ones. This has come from a combination of the incorporation of new measurements, refinements of the underlying physics, and improvements to satellite geometry modeling (Emmert, 2015; March et al., 2019; Mehta et al., 2017). A major obstacle in enhancing our understanding of the thermosphere is the scarcity of in situ measurements, particularly in low altitude regions

(Palmroth et al., 2021). Imperfect geometry and gas-surface interaction modeling currently contributes to inconsistencies between current in situ density data sets (March et al., 2021).

The thermosphere is the neutral region of the upper atmosphere ranging from ~90 km to 500–1,000 km depending on space weather conditions. The thermosphere is influenced by complex internal and external factors. Extreme Ultraviolet (EUV) emissions from the Sun heat up the thermosphere forcing expansion and increasing density at a given location. Effects from solar emissions can be represented by different solar indices and proxies (e.g., $F_{10.7}$), which serve as reliable model drivers (Tobiska et al., 2008). The Sun continuously emits particles in the form of solar wind, which interact with and disrupt the Earth's magnetosphere. During eruptive events, such as coronal mass ejections, there are strong interactions between the particles and the magnetosphere that result in energy being deposited into the thermosphere through Joule heating and particle precipitation at high latitudes (Knipp et al., 2004). This energy enhancement causes large increases in density and is difficult to model (Bruinsma et al., 2021; Oliveira et al., 2017). The planetary amplitude index (*ap*) has 28 discrete values that represent the level of global geomagnetic activity. During geomagnetic storms, the storm-time disturbance index (*Dst*) can be used to improve density modeling (Bowman et al., 2008).

These modeling limitations – both driver forecasting and underlying model uncertainty – put a stress on Space Domain Awareness (SDA), which shifts the focus from catalog maintenance to threat assessment (Holzinger & Jah, 2018). In SDA, decisions are made based on uncertainty information, when available. Uncertainties in density modeling can result in large discrepancies between expected and observed satellite positions (Bussy-Virat et al., 2018; R. Licata et al., 2020, 2021). The most recent Drag Temperature Model (DTM-2020) became the first prominent thermosphere model to provide uncertainty estimates (Bruinsma & Boniface, 2021). They determined 1-σ uncertainties as a function of the inputs (Boniface & Bruinsma, 2021). Another major improvement in density modeling capabilities came with the introduction of real-time data assimilation. The High Accuracy Satellite Drag Model (HASDM; Storz et al., 2005) is an assimilative model that leverages an empirical model (JB2008) and Dynamic Calibration of the Atmosphere (DCA) to correct the density nowcast with satellite observations. The updated nowcast can then be propagated forward in time. This model/technique arguably represents the current state of the art in our ability to accurately predict atmospheric density. HASDM outputs had not been publicly available until the recent release of the SET HASDM density database (Tobiska et al., 2021).

The continued commercial satellite launches are creating a necessity for managing the increasing traffic in orbit. A model that balances prediction accuracy and robustness of uncertainty estimates will be a critical tool for the rising importance of Space Traffic Management (Muelhaupt et al., 2019). In this work, we investigate different modeling techniques using machine learning (ML) to create a surrogate for the database. This allows us to extrapolate beyond the limits of the existing database. A straightforward way to accomplish this is by creating a nonlinear regression model on the full density grid. However, this creates limitations in the context of uncertainty quantification (UQ). To combat this, we focus on the development of nonlinear reduced order models (ROMs). For dimensionality reduction, a linear technique called principal component analysis (PCA) is utilized. PCA has been applied to this HASDM data set by Licata, Mehta, Tobiska et al. (2021) for scientific investigation but is used here as part of the modeling framework. The predictive models are standard feed-forward neural networks or artificial neural networks (ANNs) that employ nonlinear activation functions. We explore various custom loss functions in training in order to obtain reliable or calibrated uncertainty estimates that are crucial for satellite collision avoidance operations.

The paper is organized as follows. We first explain the data (mass density outputs from the HASDM database) and the various techniques used for model development. For clarification, the density space refers to the full 3D density grid (or full state), while the latent space refers to the PCA coefficients (or reduced state). The ML models operate in the latent space. Then, we show the performance and UQ capabilities using a standard mean square error (MSE) loss function. We present a method to obtain calibrated uncertainty estimates of the ML model outputs using the negative logarithm of predictive density (NLPD) and the continuous ranked probability score (CRPS) loss functions, which are based on the probability and cumulative distribution functions (CDFs) of the normal distribution. The best model (called HASDM-ML), in terms of accuracy and reliability, is then compared to the HASDM database in the density space.

2. Methodology

2.1. Data

HASDM is an assimilative framework using the Jacchia-Bowman 2008 Empirical Thermospheric Density Model (JB2008; Bowman et al., 2008) as a background density model. HASDM improves upon the density correction work of Marcos et al. (1998) and Nazerenko et al. (1998) to modify 13 global temperature correction coefficients with its DCA algorithm. HASDM uses observations of more than 70 carefully chosen calibration satellites to estimate local density values. The satellite orbits span an altitude range of 190–900 km although a majority are between 300 and 600 km (Bowman & Storz, 2003). The HASDM algorithm uses a prediction filter that employs wavelet and Fourier analysis for the correction coefficients (Storz et al., 2005). Another highlight of HASDM's novel framework is its segmented solution for the ballistic coefficient. This allows the ballistic coefficient estimate to deviate over the fitting period for the satellite trajectory estimation.

SET validates the HASDM output each week and archives the results. The archived values from 2000 to 2020 make up the SET HASDM density database, upon which this work is based. The database contains density outputs with 15° longitude, 10° latitude, and 25 km altitude increments spanning from 175 to 825 km. This results in 12,312 grid points for every three hours from the start of 2000 to the end of 2019. For further details on HASDM, the reader is referred to Storz et al. (2005) and for details on SET's validation process and on the database the reader is referred to Tobiska et al. (2021).

Solar and geomagnetic indices/proxies are the key drivers for JB2008, and subsequently HASDM. $F_{10.7}$ (referred to as F_{10} in this work) is a solar proxy used widely in thermospheric density models. It represents 10.7 cm solar radio flux which does not directly interact with the thermosphere, but it is a reliable proxy for solar EUV heating. The S_{10} index is a measure of the integrated 26–34 nm solar EUV emission, which is absorbed by atomic oxygen in the middle thermosphere. M_{10} is a proxy that represents the Mg II core-to-wing ratio and is a surrogate for far ultraviolet photospheric 160-nm Schumann-Runge Continuum emissions that cause molecular oxygen dissociation in the lower thermosphere. The last solar index used by JB2008 is Y_{10} which is a hybrid measure of solar coronal X-ray emissions and Lyman-alpha. S_{10}, M_{10}, and Y_{10} have no relation to the 10.7 cm wavelength but are converted to solar flux units (sfu) through linear regression with F_{10}. The 81-day centered averages of these four drivers are used as inputs to JB2008 and are denoted by the "*81c*" subscript, resulting in a total of eight drivers for solar activity.

For temperature equations corresponding to geomagnetic activity, JB2008 utilizes *ap* and *Dst*. The 3-hr *ap* is indicative of overall geomagnetic activity and is only measured at low to mid-latitudes. It takes on 1 of 28 values on an uneven discrete grid between 0 and 400. JB2008 uses *Dst* during geomagnetic storms if the minimum $Dst < -75\ nT$, and it is the primary parameter for energy being deposited into the thermosphere during storms. For more information on the JB2008 model, the reader is referred to Bowman et al. (2008). Information on how to access and use JB2008 can be found at the link in the Data Statement. The data are split into training, validation, and test sets using 60%, 20%, and 20%, respectively, as shown in Figure 1. The data splitting scheme was chosen to include a full range of conditions in the training set while leaving enough diverse data for the validation and independent test set. Table 1 shows the number of time steps in the SET HASDM density database across various space weather conditions. The cutoff values for F_{10} and *ap* are obtained from Licata, Tobiska, and Mehta (2020) where they were used to bin space weather driver forecasts.

In Table 1, there is clear under representation of geomagnetic storms in this vast data set. This can cause limitations in model development, because over 98% of the data set corresponds to $ap \leq 50$. Hierarchical modeling could be used for data of this nature, but we proceed with the development of a single comprehensive model. This decision was made due to the limited data for high geomagnetic activity and for simplicity in model development.

2.2. Principal Component Analysis

PCA is an eigendecomposition technique that determines uncorrelated linear combinations of the data that maximize variance (Hotelling, 1933; Pearson, 1901). PCA has been previously used to analyze atmospheric density models and accelerometer-derived density sets. Some researchers have applied PCA to CHAllenging Minisatellite Payload (CHAMP) and Gravity Recovery and Climate Experiment (GRACE) accelerometer-derived density data in order to analyze the dominant modes of variation and identify phenomena encountered by the satellites

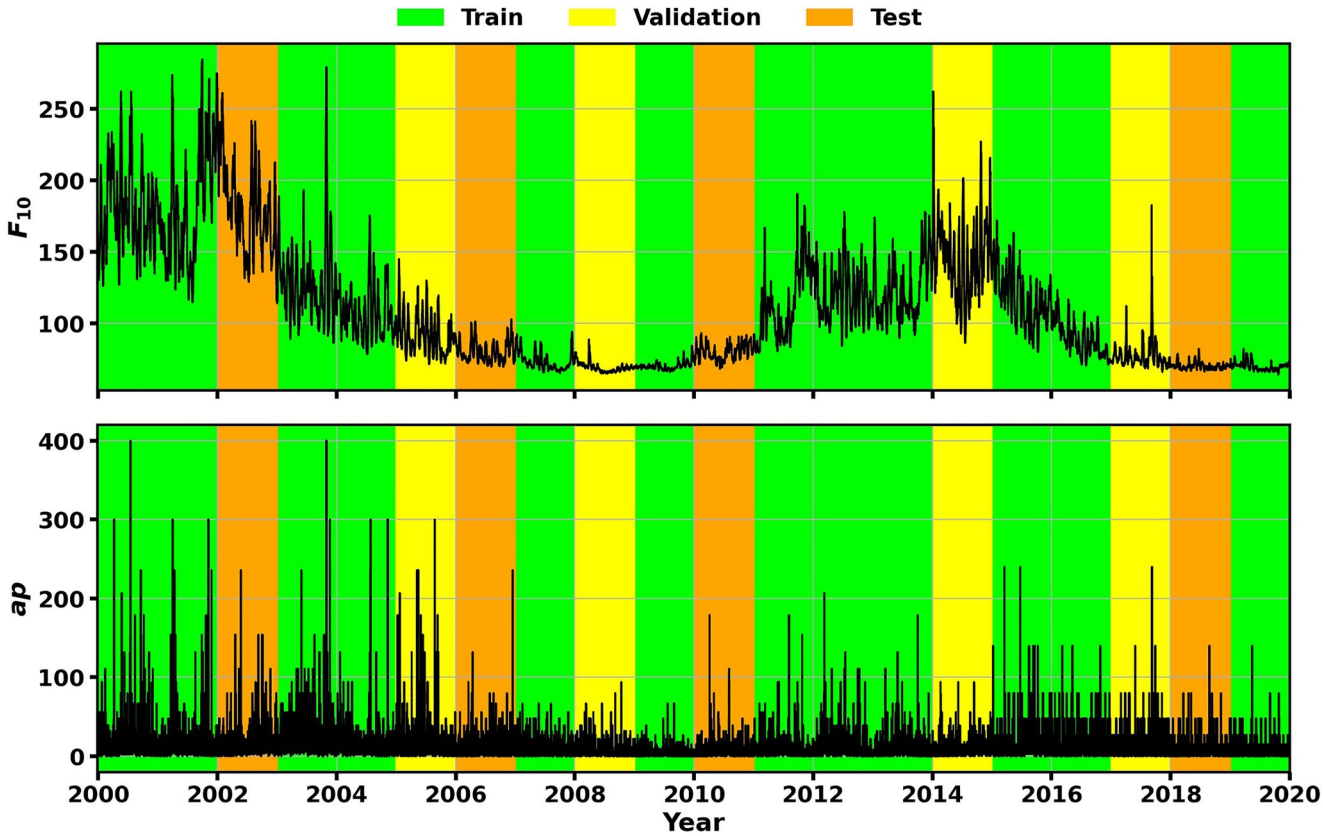

Figure 1. F_{10} and planetary amplitude index at available data points shaded to show the training, validation, and test splits.

(Calabia & Jin, 2016; Lei et al., 2012; Matsuo & Forbes, 2010). PCA has also been leveraged for dimensionality reduction to create linear dynamic ROMs (Gondelach & Linares, 2020; Mehta and Linares, 2017, 2020; Mehta et al., 2018). Licata, Mehta, Tobiska et al. (2021) recently applied PCA to the orbit-position-derived SET HASDM density database to identify the dominant modes of variability and to compare the database to JB2008.

The SET HASDM data set is a prime candidate for PCA with the goal of predictive modeling due to its high dimensionality; the 12,312 HASDM grid locations would inhibit further UQ work due to computer memory constraints. The three spatial dimensions are first flattened – reshaped into a column of the 12,312 components– to make the data set two dimensional (time × space). Then a common logarithm (log_{10}) of the density values is taken in order to reduce the variance of the data set. Next, we subtract the temporal mean over all 20 years for each cell to center the data and save it for later use. Finally, we perform PCA using the *svds* function in *MATLAB*. PCA decomposes the data and separates spatial and temporal variations as in Equation 1,

Table 1
Number of Time Steps for Different Space Weather Conditions Across the Space Environment Technologies High Accuracy Satellite Drag Model Density Database

	$F_{10} \leq 75$	$75 < F_{10} \leq 150$	$150 < F_{10} \leq 190$	$F_{10} > 190$	All F_{10}
ap ≤ 10	13,839	22,034	4,126	2,088	42,087
10 < ap ≤ 50	3,003	9,226	1,982	1,091	15,302
ap > 50	54	652	196	149	1,051
All ap	16,896	31,912	6,304	3,328	58,440

$$\mathbf{x}(\mathbf{s},t) = \bar{\mathbf{x}}(\mathbf{s}) + \tilde{\mathbf{x}}(\mathbf{s},t) \quad \text{and} \quad \tilde{\mathbf{x}}(\mathbf{s},t) = \sum_{i=1}^{r} \alpha_i(t)\, U_i(s) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the log-transformed model output state (HASDM density at all $n = 12{,}312$ grid locations), $\bar{\mathbf{x}}$ is the mean, $\tilde{\mathbf{x}}$ is the deviation from the mean, $\mathbf{s}$ represents the spatial dimension, r is the choice of order truncation (here $r = 10$), $\alpha_i(t)$ are temporal coefficients, and U_i are orthogonal modes or basis functions. The value of r is chosen, because it captures approximately 90% of the variance and has been shown to be an optimal choice for observability and accuracy for data assimilation (Mehta & Linares, 2017; Mehta et al., 2018). Equation 2 shows how to derive these modes.

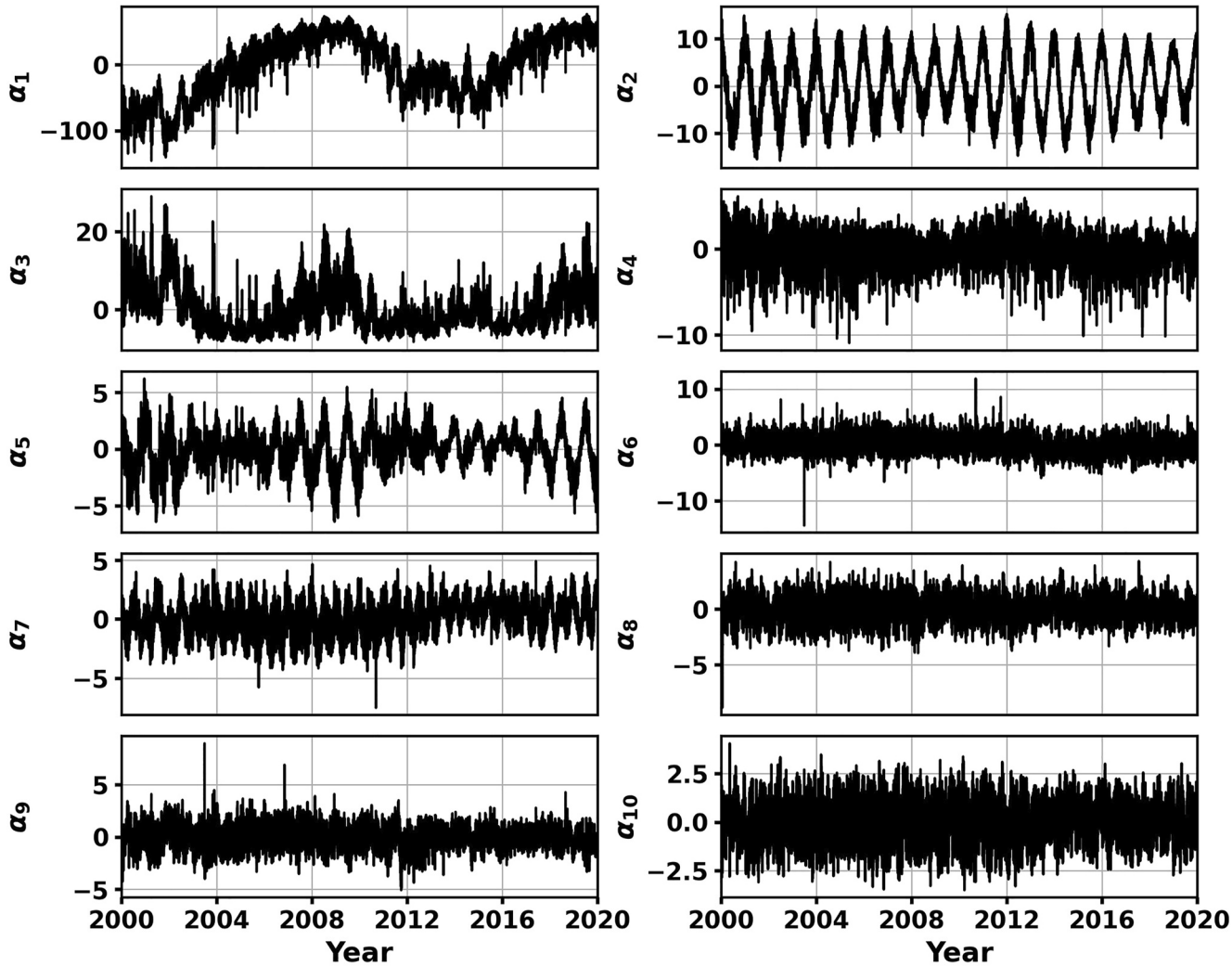

Figure 2. First 10 principal component analysis coefficients (α_i) for the High Accuracy Satellite Drag Model database.

$$\mathbf{X} = \begin{bmatrix} | & | & | & & | \\ \widetilde{\mathbf{x}}_1 & \widetilde{\mathbf{x}}_2 & \widetilde{\mathbf{x}}_3 & \dots & \widetilde{\mathbf{x}}_m \\ | & | & | & & | \end{bmatrix} \text{ and } \mathbf{X} = U\Sigma V^T \tag{2}$$

In Equation 2, m represents the ensemble size (two solar cycles with HASDM or m = 58,440 time steps) and $\mathbf{X}$ represents the log-transformed, mean-centered data with column $\widetilde{\mathbf{x}}_1 = \widetilde{\mathbf{x}}(\mathbf{s}, 1)$ in Equation 1. The left unitary matrix, U, is made of orthogonal vectors that represent the modes of variation, Σ is a diagonal matrix consisting of the squares of the eigenvalues that correspond to the vectors in U, and V is a matrix composed of the right singular vectors of $\mathbf{X}$. We encode the data ($\mathbf{X}$) into temporal coefficients (α_i) by performing matrix multiplication between Σ and V^T. To decode back to the density, the coefficients are multiplied by U with the temporal mean added at each cell. Taking the antilogarithm (10^y) of the resulting values completes the back transformation. Figure 2 shows the first 10 PCA coefficients generated using the methods described above which will be used to train the model (described in Sections 2.4 and 2.5). The first five PCA coefficients are discussed in detail by Licata, Mehta, Tobiska et al. (2021). They noted that there was strong correlation between α_1 and solar activity and between α_2 and annual variations. Beyond that the coefficients were correlated with different space weather or temporal inputs with various strength depending on the phase of the solar cycle.

Table 2
List of Inputs in the Two Sets Used for Model Development

JB2008 Inputs			Historical JB2008 Inputs		
Solar	Geomagnetic	Temporal	Solar	Geomagnetic	Temporal
F_{10}, S_{10}, M_{10}, Y_{10}, F_{81c}, S_{81c}, M_{81c}, Y_{81c}	ap, Dst	t_1, t_2, t_3, t_4	F_{10}, S_{10}, M_{10}, Y_{10}, F_{81c}, S_{81c}, M_{81c}, Y_{81c}	ap_A, ap, ap_3, ap_6, ap_9, ap_{12-33}, ap_{36-57}, Dst_A, Dst, Dst_3, Dst_6, Dst_9, Dst_{12}, Dst_{15}, Dst_{18}, Dst_{21}	t_1, t_2, t_3, t_4

2.3. Hyperparameter Tuning

In ML, developing an accurate model previously required a manual architecture selection process that did not guarantee optimal performance for a given application (Elsken et al., 2019). Recent developments in hyperparameter tuning make this process much quicker and more thorough as there is significantly less user intervention, and hence a saving in time needed to run this process. In this work, we use KerasTuner which allows the user to define the ranges and choices of any hyperparameter and to choose a search algorithm (Abdelminaam et al., 2021; Haegel & Husa, 2020; Licata, Mehta, Weimer et al., 2021; O'Malley et al., 2019; Rogachev & Melikhova, 2020). All subsequent mentions to a tuner refer to the use of the KerasTuner package. We run all tuners for 100 trials (or architectures) with the first 25 being randomly sampled from our hyperparameter space while the last 75 trials take advantage of the tuner's Bayesian optimization scheme. We allow the tuner to choose an optimizer, the number of hidden layers, the number of neurons in each hidden layer, and activation functions. The tuner can choose unique neuron counts and activation functions within each layer. The number of total trials and initial/random trials was chosen through early experimentation. We found that increasing the number of trials simply increased the model development time without any noteworthy performance improvements.

The tuner trained three identical models for each trial (repeat weight initialization and training) and returned the model with the lowest validation loss after 100 training iterations or epochs. The Bayesian optimization scheme chooses future architectures based on the validation losses resulting from previous trials. Once completed, the tuner returns the best 10 models, which we can evaluate on the training and validation sets to find the most accurate and generalized model. The metrics used to evaluate the resulting models are mean absolute percent error (for accuracy) and the calibration error score (see Section 2.5.1).

2.4. Regression Modeling

A traditional approach to regression modeling with UQ is to develop Gaussian process (GP) models (Chandorkar et al., 2017; Rasmussen, 2004). In the early stages of model development, we attempted this approach – training GP regression models on each of the PCA coefficients. However, we could only use 10 years of data for training, limited by computational resources or random access memory, which resulted in higher predictive error. In addition, the resulting models (10 GP regression models each trained on a single PCA coefficient) were 6.83 GB each. This makes subsequent work cumbersome as we would need to load 68.3 GB of models to make evaluations. Therefore, the results in this work only pertain to the feedforward neural networks with MC dropout.

There are three methods we explore in developing HASDM-ML. First, we create a ROM using PCA for dimensionality reduction and a nonlinear ANN for prediction with the MSE loss function. Note: the term "prediction" here refers to model outputs, not forecasts. We then modify this technique using a custom loss function in an attempt to obtain a model with calibrated uncertainty estimates. We test two loss functions to achieve this, NLPD and CRPS; details of which are given in the following section. Loss functions inform the model of the objective, so the weights can be modified to minimize the losses. ANNs have two key components: features (or inputs) and labels (or outputs). We try using three separate input sets for our regression models, the first two are explained in Table 2. The first set is the JB2008 input set, referred to as JB. Licata, Mehta, Tobiska et al. (2021) showed that the SET HASDM density database contained evidence of poststorm cooling mechanisms which cannot be captured solely by geomagnetic indices at epoch. Therefore, we introduce a second set that is similar to the first

but with a time history of the geomagnetic indices. We refer to this as JB_H (Historical JB2008). Unlike the actual JB2008 inputs, all of our input sets contain sinusoidal transformations to the day of year (doy) and universal time (UT) inputs. The four resulting temporal inputs (shown in Equation 3) represent annual (t_1 and t_2) and diurnal (t_3 and t_4) variations. The generic form ($2\pi x/y$) allows for the input (x) to oscillate between -1 and 1 over the period (y). The use of cyclical functions of time as inputs has been previously demonstrated (Licata, Mehta, Weimer et al., 2021; Weimer et al., 2020).

$$t_1 = \sin\left(\frac{2\pi doy}{365.25}\right), \quad t_2 = \cos\left(\frac{2\pi doy}{365.25}\right), \quad t_3 = \sin\left(\frac{2\pi UT}{24}\right), \quad t_4 = \cos\left(\frac{2\pi UT}{24}\right). \tag{3}$$

In the JB_H set, the geomagnetic indices are extensive in an effort to improve storm-time and poststorm modeling. The "A" subscript (seen in the geomagnetic column of Table 2) refers to the daily average and the numerical subscripts refer to the value of the index that many hours prior to the epoch. The combination of two numbers references the number of previous hours the index is being averaged over (e.g., $ap_{12\text{-}33}$ refers to the average *ap* value between 12 and 33 hr prior to the prediction epoch). The *ap* time series is the same one used in the Naval Research Laboratory Mass Spectrometer Incoherent Scatter Radar Extended model (NRLMSISE-00; Picone et al., 2002). The authors found that using different time histories of *ap* and *Dst* (shown in Table 2) resulted in generally more calibrated models on independent data (see Section 3). For completeness, the results will also be shown using an input set that adopts the same time history for *Dst* as the *ap* time history in Table 2, both geomagnetic indices using the NRLMSISE-00 time series. This input set will be referred to as JB_{H0}.

2.5. Uncertainty Quantification

Dropout is a regularization tool often used in ML to prevent the model from overfitting to the training data (Srivastava et al., 2014). In standard feed-forward neural networks, each layer sends outputs from all nodes to those in the subsequent layer where they are introduced to weights and biases. Deep neural networks can have millions of parameters and thus are prone to overfitting. This causes undesired performance when interpolating or extrapolating.

Dropout layers use Bernoulli distributions (a binomial distribution with one trial), one for each node, with probability P. This makes the model probabilistic since the distributions are sampled each time that a set of inputs are given to the model. If a sampled Bernoulli is a "1", the node's output is nullified and the output of the layer is scaled based on the number of nullified outputs. Dropout is believed to make each node independently sufficient and not reliant on the outputs of other nodes in the layer (Alake, 2020).

In traditional use, dropout layers are only activated during training to uphold the deterministic nature of the model. However, measures can be taken in order for this feature to remain activated during prediction making the model probabilistic. When passing the same input set to the model a significant number of times (e.g., 1,000), there is a distribution of model outputs for each unique input. This process is referred to as Monte Carlo (MC) dropout. Essentially, every time the model is presented with a set of inputs, random nodes are dropped out providing a different functional representation of the model. Gal and Ghahramani (2016) show that MC dropout is a Bayesian approximation of a GP.

In this case, the model uses the input set to predict the 10 PCA coefficients. Using the MC samples, we estimate the sample mean and variance for each of the predicted coefficients/outputs (Efron & Tibshirani, 1993). The loss is computed for each output separately. Each unique input can be passed to the model k times and there will be $k \times 10$ outputs. The mean and variance are computed from those outputs with respect to the repeated axis, k. The two loss functions used to improve uncertainty estimation (in addition to MSE) are NLPD and CRPS. NLPD is based on the logarithm of the probability distribution function (pdf) of the Gaussian distribution, and is shown in Equation 4 (Camporeale & Care, 2021; Matheson & Winkler, 1976).

$$NLPD(y, \mu, \sigma) = \frac{(y-\mu)^2}{2\sigma^2} + \frac{\log\left(\sigma^2\right)}{2} + \frac{\log(2\pi)}{2} \tag{4}$$

In Equation 4, y is the ground truth (α_i), μ is the sample mean of the prediction, and σ is the sample standard deviation of the prediction, each being computed for all 10 outputs. For clarity, the *log* used in the NLPD loss function is the natural logarithm. The second loss function for uncertainty estimation is CRPS which is shown

analytically in Equation 5 (Gneiting et al., 2005). The main difference between NLPD and CRPS is that CRPS is also based on the CDF of the Gaussian distribution as opposed to only the pdf.

$$CRPS(y,\mu,\sigma) = \sigma\left[\frac{y-\mu}{\sigma}\mathrm{erf}\left(\frac{y-\mu}{\sqrt{2}\sigma}\right) + \sqrt{\frac{2}{\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) - \frac{1}{\sqrt{\pi}}\right]. \tag{5}$$

An important aspect of using the loss functions described in Equations 4 and 5 is the preparation of the training data. The data are traditionally in the following form. The features are set up as the number of samples (n), with n_I denoting the number of inputs, resulting in the input shape ($n \times n_I$). The labels are set up as the number of samples with n_o being the number of outputs, resulting in the output shape ($n \times n_o$). To implement these loss functions, we stack each input and output by the number of MC samples, k. This is a repeated axis, meaning all samples along this axis are identical about k; the samples are not identical about n. The resulting shapes of the features and labels are ($n \times k \times n_I$) and ($n \times k \times n_o$), respectively. This allows us to determine the mean and standard deviation for each sample in the batch within the loss function.

2.5.1. Latent Space UQ

Since there are multiple models and loss functions to compare, we have to implement a metric to judge each model's ability to provide reliable uncertainty estimates. To do so, we modified a calibration error equation from Equation 4 in Anderson et al. (2020) shown as

$$\text{Calibration Error} = \frac{100\%}{r \cdot m}\sum_{i=1}^{r}\sum_{j=1}^{m}\left|p\left(\alpha_{i,j}\right) - p\left(\hat{\alpha}_{i,j}\right)\right| \tag{6}$$

In Equation 6, m is the number of prediction intervals investigated, r is the choice order of truncation for PCA (the number of model outputs), $p(\alpha)$ is the expected cumulative probability (or prediction interval), and p($\hat{\alpha}$) is the observed cumulative probability. The prediction intervals of interest in this work span from 5% to 99% with 5% increments – (0.05, 0.10, 0.15, …, 0.90, 0.95, 0.99). P($\hat{\alpha}$) is computed empirically shown in Equation 7,

$$p\left(\hat{\alpha}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\hat{\alpha}_i^l < \alpha_i < \hat{\alpha}_i^u\right) \tag{7}$$

where n is the number of samples, $\mathbb{I}$ is the indicator function, $\hat{\alpha}_i^l$ is the lower bound of the prediction interval, $\hat{\alpha}_i^u$ is the upper bound of the prediction interval, and α_i is the sample. The indicator function returns a one if the inequality is true and a 0 otherwise. To compute the bounds, we use the pair of equations given in Equation 8 (Davison et al., 1997; Pevec & Kononenko, 2014).

$$\hat{\alpha}_i^l = \mu - z\sigma \;\text{ and }\; \hat{\alpha}_i^u = \mu + z\sigma, \tag{8}$$

where z is the critical value used for the prediction interval. This is calculated using the Gaussian CDF

$$z = \sqrt{2}\,\mathrm{erf}^{-1}\left(PI\right) \tag{9}$$

where PI is the prediction interval of interest (e.g., 95% or 0.95). Comparing the expected ($p(\alpha)$) and observed ($p\left(\hat{\alpha}\right)$), cumulative probabilities is done qualitatively by plotting the calibration curves: $p\left(\hat{\alpha}\right)$ versus $p(\alpha)$. The curves show how well calibrated the uncertainty estimates are at capturing the appropriate percentage of true samples. A perfectly calibrated model will have a straight 45° calibration curve. If a calibration curve is above or below the 45° reference line, the model is over or underestimating the uncertainty, respectively. The calibration error score (Equation 6) is a quantitative measure of the average deviation from perfect calibration in the latent space, averaged across each output. In this work, we measure robustness and reliability through the calibration error score and the calibration curves. Since this refers to latent space calibration, we use α in Equations 6–9 but for density (Section 2.5.2), α would be replaced with ρ.

2.5.2. Density UQ

While latent space calibration is important because the model is trained on PCA coefficients, determining the reliability of the model's predictions on the resulting density is the ultimate goal. To examine this, we look at the orbits of CHAMP (Luhr et al., 2002) and GRACE (Bettadpur, 2012). These satellites had onboard accelerometers

Table 3
Information on the Four Storms Used in the Calibration Analysis

Start Data	F_{10} (Min – Max)	Max *ap*	Min *Dst*	Set
21 May 2002	180.3–189.1	236	−109	Test
30 September 2002	135.8–161.7	154	−176	Test
28 October 2003	166.9–279.1	400	−383	Training
7 November 2004	94.9–140.9	300	−373	Training

from which mass density has been estimated (Bruinsma & Biancale, 2003; Calabia & Jin, 2016; Doornbos, 2012; Liu et al., 2005; March et al., 2019; Mehta et al., 2017; Sutton, 2008). Licata, Mehta, Tobiska et al. (2021) recently compared the SET HASDM density database and JB2008 to both CHAMP and GRACE-A density estimates over the lifetime of their measurements. We use the satellite positions for density calibration assessment between HASDM and HASDM-ML, because it is computationally inefficient to compute the calibration curve across all 719, 513, 280 density values in the HASDM database (58,440 × 12,312). Performing that assessment would also be difficult in terms of visualization.

We evaluate HASDM-ML 1,000 times every 3 hr across the entire availability of CHAMP (2002–2010) and GRACE (2002–2011) position data listed in the measurements presented by Mehta et al. (2017) and interpolate them to the satellite locations using trilinear interpolation. For clarification, only the satellite positions are considered, not the density estimates. This model evaluation and interpolation allows us to compute the observed cumulative probability of HASDM-ML relative to the HASDM database in terms of density. Due to the redundancy and computational expense, we only interpolate the model and database density every 500 samples (5,000 and 2,500 s for CHAMP and GRACE-A, respectively). The CHAMP orbit comparison uses 23,795 HASDM prediction epochs (40.7% of the total available HASDM data) with the density being interpolated to at least two satellite positions per prediction due to the cadence of this study. The GRACE orbit comparison uses 24,602 HASDM prediction epochs (42.1% of the total available HASDM data) with the density being interpolated to at least four satellite positions per prediction. The number of satellite positions per prediction comes from the number of positions used every 3 hours (HASDM cadence). This provides a wide view of the model's UQ capabilities considering the wide array of positions and conditions covered. To perform these simulations, the model had to be evaluated 23, 795, 000 and 24, 602, 000 times for CHAMP and GRACE, respectively. These numbers come from the number of HASDM prediction epochs and the number of MC runs (1,000). HASDM-ML can perform these predictions in 17.27 s for CHAMP and 17.54 s for GRACE on a laptop with a NVIDIA GeForce GTX 1070 Mobile graphics card. Using the central processing unit, the model takes 143 s for CHAMP and 152 s for GRACE.

Geomagnetic storms are particularly difficult conditions to model accurately. Therefore, we look at four geomagnetic storms from 2002 to 2004 where we can evaluate HASDM-ML's reliability across unique events. Two of the events take place in 2002, which is outside the training data set, while the other two are from 2003 to 2004 and are seen in training. Information on these storms can be found in Table 3. To conduct this assessment, we employ the methodology from the Licata, Mehta, Tobiska et al. (2021) storm-time case study. The model is evaluated over a 6-day period encompassing a storm then interpolated to the CHAMP locations (10 s cadence). Again, the interpolation to satellite positions is conducted to assess and visualize the implications of density UQ along a satellite orbit. During each three-hour prediction period, the density grids remain constant. All 1,000 HASDM-ML density variations are then averaged across each orbit. We consider the average altitude for each 6-day period to estimate the orbital period. The mean and 95% prediction interval bounds are computed to compare to the corresponding HASDM densities and shown in Figure 7 in an orbit-averaged form. We also show the orbit-averaged JB2008 predictions for comparison. In total, the six days amount to 48 model prediction epochs that result in 51,840 interpolated densities (1,000 MC runs) from which we compute the observed cumulative probabilities.

3. Results and Discussion

Upon running each input set with all three loss functions through individual tuners, the best 10 models from each tuner (in terms of training and validation metrics) are evaluated on the entire training, validation, and test sets 1,000 times. The mean absolute error results from the best of the 10 models for each input set/loss function are shown in Table 4. The mean absolute error is computed for the model prediction (mean coefficient predictions converted to density through PCA) relative to the HASDM database.

The addition of historical geomagnetic indices clearly improves the model performance with error reductions ranging from 0.72% to 2.09% (comparing the JB columns to the columns of JB_H and JB_{H0}). As mentioned in Section 2.4, the motivation for using the time series geomagnetic indices was to improve storm-time and

Table 4
Mean Absolute for the Best Model From Each Technique Across Training, Validation, and Test Data

	Training			Validation			Test		
Technique	JB	JB_H	JB_{H0}	JB	JB_H	JB_{H0}	JB	JB_H	JB_{H0}
MSE	10.38%	8.73%	8.47%	12.00%	10.48%	9.91%	11.95%	10.71%	10.51%
NLPD	10.07%	9.07%	8.81%	11.93%	10.69%	9.87%	11.41%	10.69%	10.05%
CRPS	9.67%	8.64%	8.26%	11.56%	10.55%	9.69%	11.76%	10.43%	10.69%

poststorm performance. However, Table 1 shows that these conditions account for a small subset of the data meaning the notable performance improvement with the JB_H and JB_{H0} input sets show that it likely improves the predictions across all conditions. In general, the CRPS models have the lowest error and the JB_{H0} models have the lowest error with respect to the input sets. Table 5 shows the calibration error score for the same models, this time using both the mean and standard deviation of the coefficient predictions (refer to Equation 6).

The incorporation of the custom loss functions reduces the calibration error score by an order of magnitude relative to models trained with MSE, which tend to underestimate the uncertainty. The best performing loss function, in regard to calibration, is NLPD. To choose the best overall model, we focus on the test performance as those data are completely independent from the training process. We weigh the calibration performance more heavily than the prediction error as reliable uncertainty estimates are the most valuable asset for a thermospheric density model. The JB_H + NLPD model is within 1% of the error of all better-performing models (Table 4), and it has the lowest test calibration error score with scores within 0.30% of all more calibrated models on the training and validation data. As the calibration error score is a composite of the scores from each PCA coefficient, we show the calibration curves of all coefficients on the independent test set for the best JB_H + NLPD model, in panel (b), alongside the best JB_H + MSE model, in panel (a), for comparison in Figure 3.

The calibration curve in panel (b) for all PCA coefficients roughly follows the perfectly calibrated 45° line with α_5 being the only coefficient that prominently underestimates uncertainty. However, there is minimal contribution to the full state (density) after the first few coefficients, so this should not greatly impact the resulting density. For PCA, the coefficients are ordered to capture most-to-least variance, so α_1 has significantly more impact on the reconstruction of the data compared to α_{10}, for example, In sharp contrast to the JB_H + NLPD results, panel (a) shows the model trained with the MSE loss function is not nearly calibrated, as is evident in Table 5. There is a significant underestimation of the uncertainty at all cumulative PIs because the model is not trained with any terms for its variance.

The JB_H + NLPD model shown in Figure 3 will be the focus of all subsequent analyses and will be referred to as HASDM-ML. To investigate the model's reliability on density in an operational nature, we look at the orbits of CHAMP and GRACE-A each over 8-year periods with a cumulative altitude range of 300–530 km. HASDM-ML was evaluated in three-hour increments from 2002 to 2011, and was interpolated to the satellite positions at all epochs discussed in Section 2.5.2. The results for the CHAMP orbit are displayed in Figure 4.

Figure 4 panel (a) shows the density ratios of HASDM-ML and JB2008 relative to HASDM. The HASDM-ML ratios have much lower variance than the JB2008 ratios. The mean ratios for both models are 1.03. However, 95% of the HASDM-ML ratios are between 0.75 and 1.25 compared to 86% for JB2008. The surrogate ML model is imperfect in its mean prediction, as seen in Table 4, but panels (b) and (d) show that the density uncertainty is

Table 5
Calibration Error Score (See Equation 6) for the Best Model From Each Technique Across Training, Validation, and Test Data

	Training			Validation			Test		
Technique	JB	JB_H	JB_{H0}	JB	JB_H	JB_{H0}	JB	JB_H	JB_{H0}
MSE	38.71%	37.44%	37.58%	39.62%	38.98%	39.01%	40.04%	39.79%	39.72%
NLPD	3.40%	3.06%	2.79%	3.08%	2.51%	2.84%	2.21%	1.76%	2.79%
CRPS	3.29%	7.93%	4.46%	2.27%	4.63%	2.73%	2.40%	2.39%	2.95%

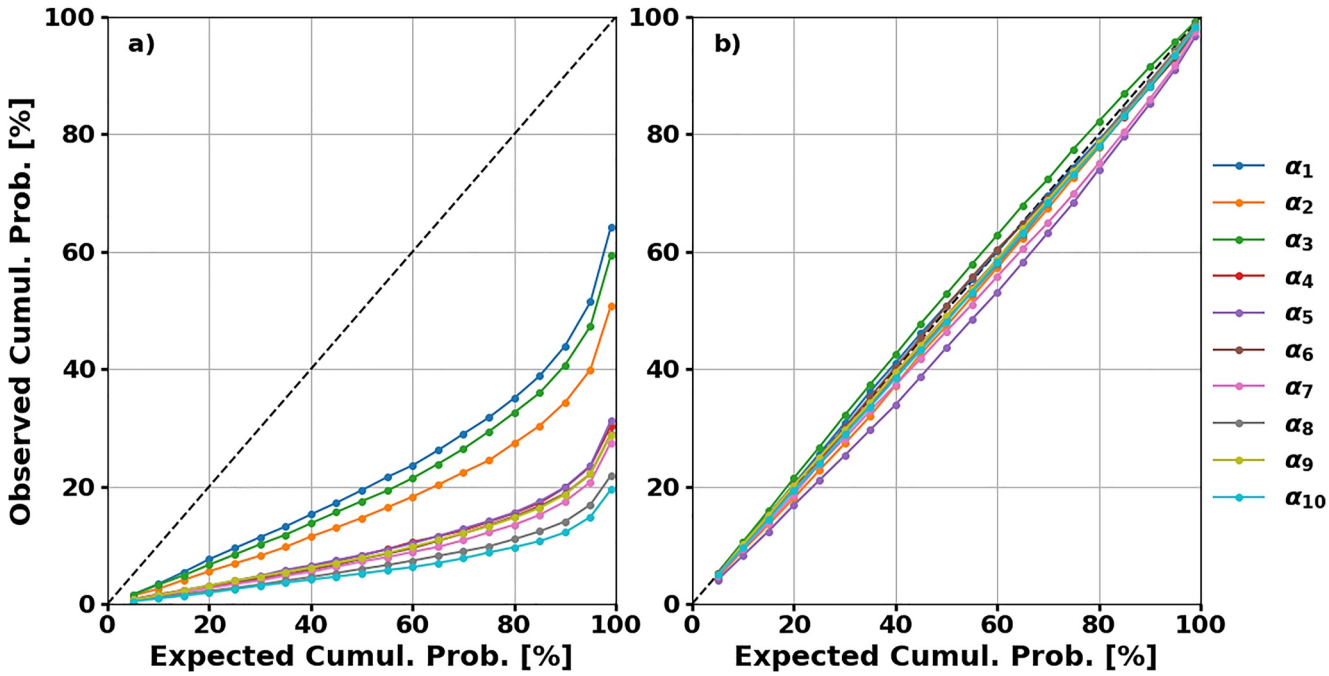

Figure 3. Expected versus observed cumulative probability of all 10 Principal Component Analysis coefficients on the test set using JB_H + MSE (a) and JB_H + NLPD (b).

reliable. The calibration curve is exceptional with the observed cumulative probability being within 1% of the expected value for all 20 cumulative PIs tested. Figure 5 shows the same analysis along the GRACE-A orbit.

Figure 5 panel (a) again shows that the HASDM-ML density ratios have much less variance than JB2008. For these GRACE-A positions, the mean density ratios are 1.05 and 1.07 for HASDM-ML and JB2008, respectively. 86% of the HASDM-ML ratios are between 0.75 and 1.25 compared to 72% of JB2008 ratios. Panels (b) and (d) also demonstrate that although the model densities are not identical to HASDM, HASDM-ML provides uncertainty estimates that are reliable. Panel (d) reveals that at higher GRACE altitudes, there is slightly less agreement with the expected and observed cumulative probabilities with the largest discrepancy being just over 1%. Scatter plots comparing HASDM-ML and JB2008 to HASDM densities for both satellite orbits are displayed in Figure 6.

Figure 6 highlights the prediction accuracy of HASDM-ML compared to JB2008. Both models are well centered on the perfect-prediction line (in black), but as seen in Figures 4 and 5 HASDM-ML has a tighter spread about this line. To clarify, this scatter plot is representative of prediction accuracy and is not the same as the calibration curves seen in other figures. The coefficient of determination (R^2) is higher for HASDM-ML along both satellite orbits, and R^2 is higher for both models along the GRACE-A orbit. Figure 7 shows HASDM and HASDM-ML orbit-averaged densities during four geomagnetic storms with prediction intervals and the associated calibration curves.

Across all of the storms investigated, the mean prediction of HASDM-ML follows the trend of HASDM density. Even when the model deviates, HASDM densities are mostly captured by the uncertainty bounds (computed using Equation 8). Panels (a) and (b) represent storms not contained in the training data set, which show that HASDM-ML is well generalized, even during these highly nonlinear events. In panel (a), HASDM-ML and JB2008 overestimate the peak density, but HASDM-ML is able to better capture the timing. For this storm, JB2008 predicts a delayed and longer impact of the geomagnetic storm. The mean absolute error for HASDM-ML and JB2008 relative to HASDM are 11.91% and 13.03%, respectively. In panel (b), both models have similar predictions to HASDM for the first peak (day 2), but JB2008 has an elongated period of density overprediction from days 4–6. The mean absolute error for HASDM-ML and JB2008 relative to HASDM for this storm are 9.86% and 14.37%, respectively. The storm in panel (c), while in the training set, highlights the improved performance of HASDM-ML. After the storm, JB2008 predicts much higher densities than both HASDM and HASDM-ML. Licata, Mehta, Tobiska et al. (2021) compared the orbit-averaged densities of HASDM and JB2008 to CHAMP

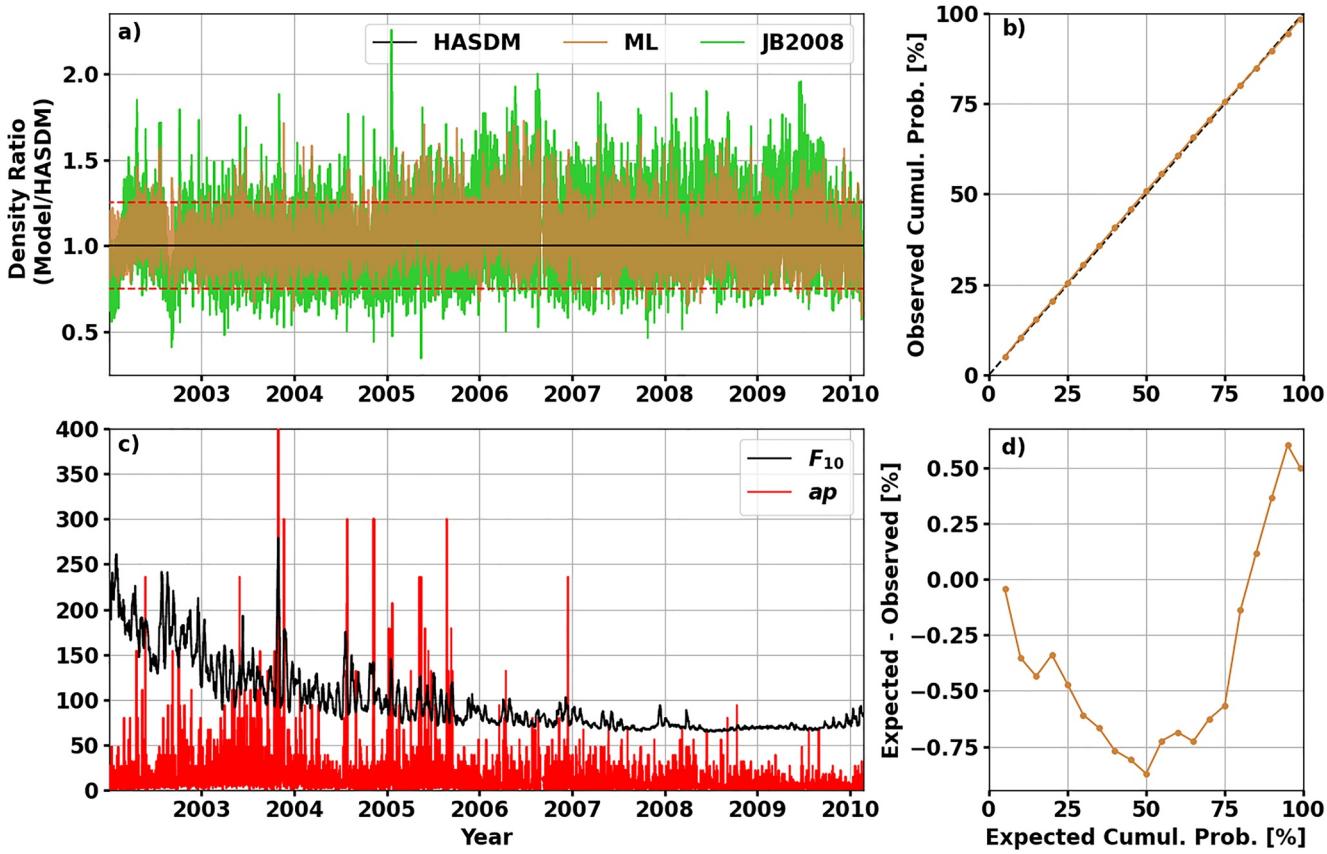

Figure 4. (a) Shows the density ratios of HASDM-ML and JB2008 relative to High Accuracy Satellite Drag Model, (b) shows the expected versus observed calibration curve, (c) shows F_{10} and *ap* for the period corresponding to (a) for reference, and (d) shows the difference between expected and observed cumulative probability corresponding to (b). Discontinuities in (a) and (c) represent data gaps. In panel (a), the red dashed lines are at ratios of 0.75 and 1.25.

and GRACE-A during this storm and found that the low poststorm densities predicted by HASDM were similar to the density estimates from both satellites which HASDM-ML is also showing. The errors for this storm are 8.46% for HASDM-ML and 25.29% for JB2008. For the last storm, panel (d), all three models predict similar trends in density. JB2008 has the most deviation, particularly between the two main phases of the storm and in the last 36 hr. The mean absolute errors are 12.64% and 19.81% for HASDM-ML and JB2008, respectively.

The calibration curves corresponding to each event show the robust nature of HASDM-ML's uncertainty estimates. None of the calibration curves, at any of 20 cumulative PIs tested, deviated from perfect calibration by more than 10.7%. We show the combination of all four calibration curves (averaged) to give a broad sense of the calibration across the storms. This curve is well calibrated and does not deviate from perfect calibration by more than 3.7%. Note: perfect calibration here is seen in the 45° line in panel (e) and the line $y = 0$ in panel (f). While the observed cumulative probability values deviated from the expected values (particularly for the individual storms), these are highly nonlinear periods where density models tend to be unreliable.

3.1. HASDM-ML Performance Metrics

To assess the conditions in which HASDM-ML can improve, the global mean absolute errors relative to HASDM are computed as a function of space weather conditions. The results are shown in Table 6 and the number of samples in each bin can be found in Table 1. The bottom half of the table contains the global mean absolute errors of JB2008 relative to HASDM for comparison.

The results from Table 6 show that HASDM-ML is robust to different F_{10} and *ap* ranges when $ap \leq 50$ since these errors do not vary by more than 2%. The only conditions in which the mean absolute error exceeds 11% is when $ap > 50$, which only accounts for 1.80% of the samples. This shows that more samples may be required for this

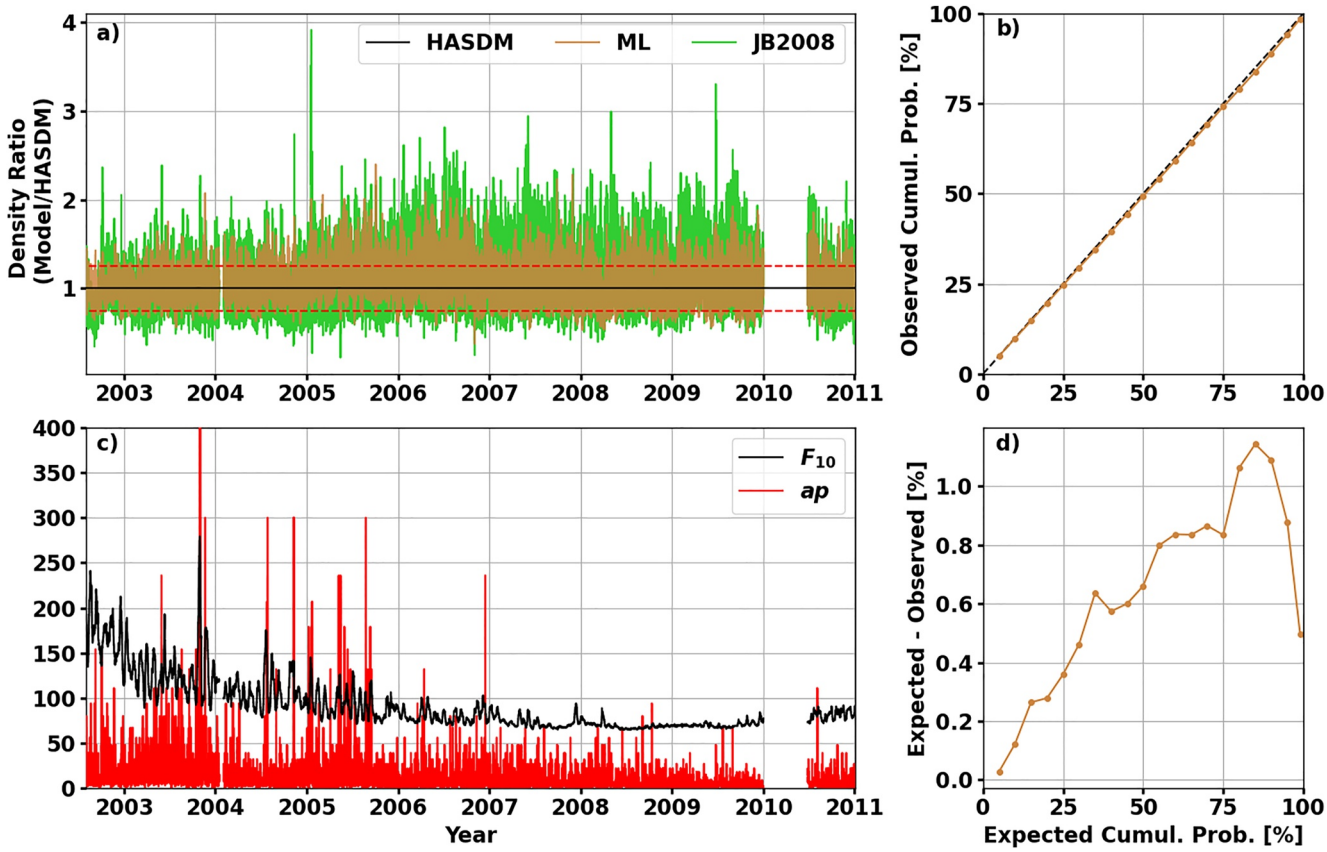

Figure 5. (a) Shows the density ratios of HASDM-ML and JB2008 relative to High Accuracy Satellite Drag Model, (b) shows the expected versus observed calibration curve, (c) shows F_{10} and *ap* for the period corresponding to (a) for reference, and (d) shows the difference between expected and observed cumulative probability corresponding to (b). Discontinuities in (a) and (c) represent data gaps. In panel (a), the red dashed lines are at ratios of 0.75 and 1.25.

specific condition in both the training and evaluation phases. The last row contains the errors only as a function of F_{10} which shows that across all four solar activity levels, the error deviates by only 1.24%. The bottom-right cell shows that the error across all 20 years of available data is only 9.71%. JB2008 densities are much less similar to HASDM. HASDM-ML is more accurate over all 20 space weather conditions considered and the improvement ranges from 3.75%–9.16%. As a function of F_{10}, HASDM-ML has the most significant improvement for low solar activity ($F_{10} \leq 75$ sfu). As a function of *ap*, HASDM-ML has the largest improvement for high geomagnetic activity ($ap > 50$). Across all the available data from the SET HASDM density database, HASDM-ML has 5.65% lower error than JB2008.

4. Summary

In this work, we developed a surrogate ML model for the SET HASDM density database with robust and reliable uncertainty estimation capabilities (Figures 4, 5 and 7). PCA was selected for dimensionality reduction due to its wide use in the field. A Bayesian search algorithm was leveraged in an attempt to identify the optimal architecture for each input set and loss function tested. We found that of the nine input-loss function combinations explored, the combination of a JB2008 input set with historical geomagnetic drivers (JB_H) and the NLPD loss function resulted in the most comprehensive model. This model, HASDM-ML, has 9.07% error across the 12-year training set and an average 10.69% error over the combined 8-year validation/test sets (Table 4). It also had the lowest calibration error score on the test set, meaning it was the most calibrated to independent data (Table 5). We also compared its calibration curves for each output across the test set to that of the MSE model with the same inputs. This showed that the MSE model considerably underestimated the uncertainty while the NLPD model was well calibrated across the 10 outputs.

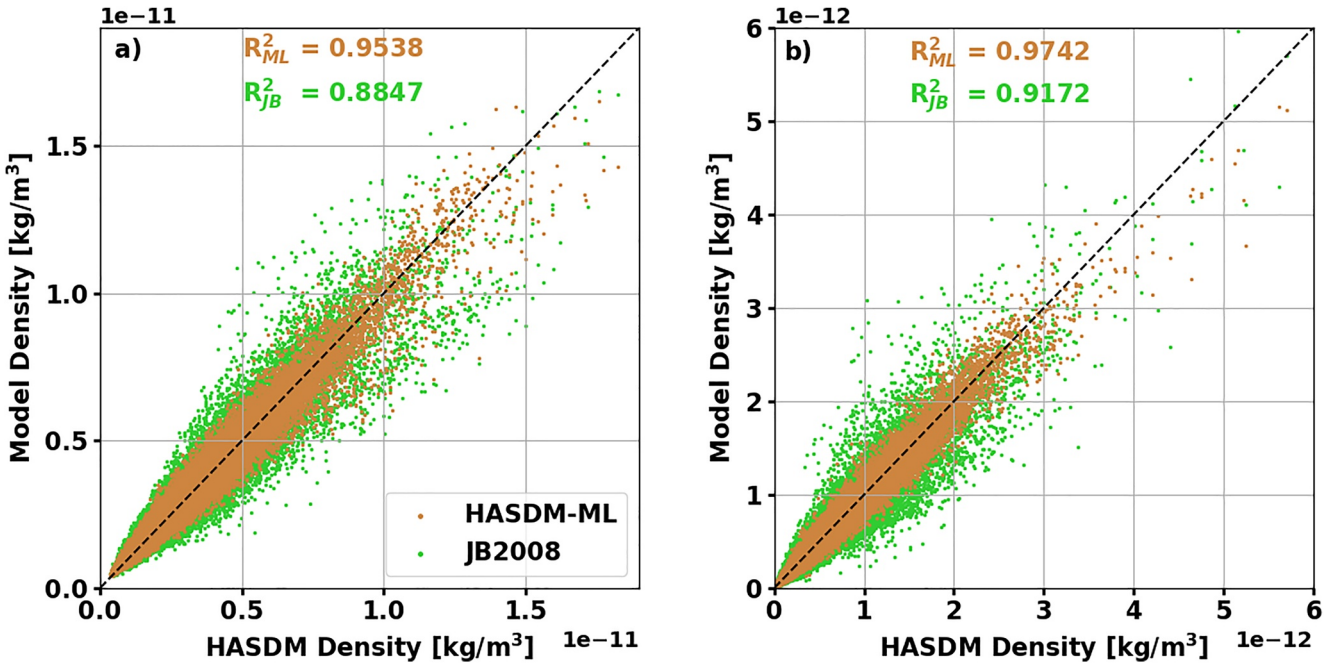

Figure 6. Scatter plot of model versus High Accuracy Satellite Drag Model (HASDM) density along the orbits of CHAllenging Minisatellite Payload (a) and Gravity Recovery and Climate Experiment-A (b). Perfect prediction would fall on the diagonal black line. The coefficient of determination (R^2) is shown for both models relative to HASDM. *Note.* ML refers to HASDM-ML while JB refers to JB2008.

Upon selecting HASDM-ML, we evaluated its uncertainty capabilities across the entire orbits of CHAMP (Figure 4) and GRACE-A (Figure 5), both using almost half of the time span of the HASDM data. This assessment showed that the mean prediction at the satellite locations closely matched that of the HASDM data set. Across all 20 prediction intervals tested over this period (2002–2011), the model provided an observed cumulative probability that never deviated more than 1% of the expected value for CHAMP's orbit and never deviated more than 1.15% for GRACE-A's orbit. A separate storm-time evaluation unveiled that across four storms, HASDM-ML provides similar density to HASDM and its uncertainty estimates remained robust and reliable (Figure 7). The results from the density calibration tests are significant, because thermospheric density models providing uncertainty estimates is still a novel concept. Additionally, uncertainty estimates themselves are not

Table 6
Mean Absolute Error Across Global Grid for HASDM-ML and JB2008 Relative to the High Accuracy Satellite Drag Model Database as a Function of Space Weather Conditions

HASDM-ML					
	$F_{10} \leq 75$	$75 < F_{10} \leq 150$	$150 < F_{10} \leq 190$	$F_{10} > 190$	All F_{10}
$ap \leq 10$	8.96%	9.78%	9.97%	9.14%	9.50%
$10 < ap \leq 50$	9.76%	10.05%	10.87%	9.90%	10.09%
$ap > 50$	15.35%	12.86%	13.23%	12.55%	13.01%
All ap	9.12%	9.92%	10.36%	9.55%	9.71%
JB2008					
	$F_{10} \leq 75$	$75 < F_{10} \leq 150$	$150 < F_{10} \leq 190$	$F_{10} > 190$	All F_{10}
$ap \leq 10$	17.42%	13.53%	14.02%	14.84%	14.92%
$10 < ap \leq 50$	17.76%	15.70%	15.17%	16.79%	16.11%
$ap > 50$	22.07%	22.77%	22.39%	18.90%	22.12%
All ap	17.49%	14.34%	14.64%	15.66%	15.36%

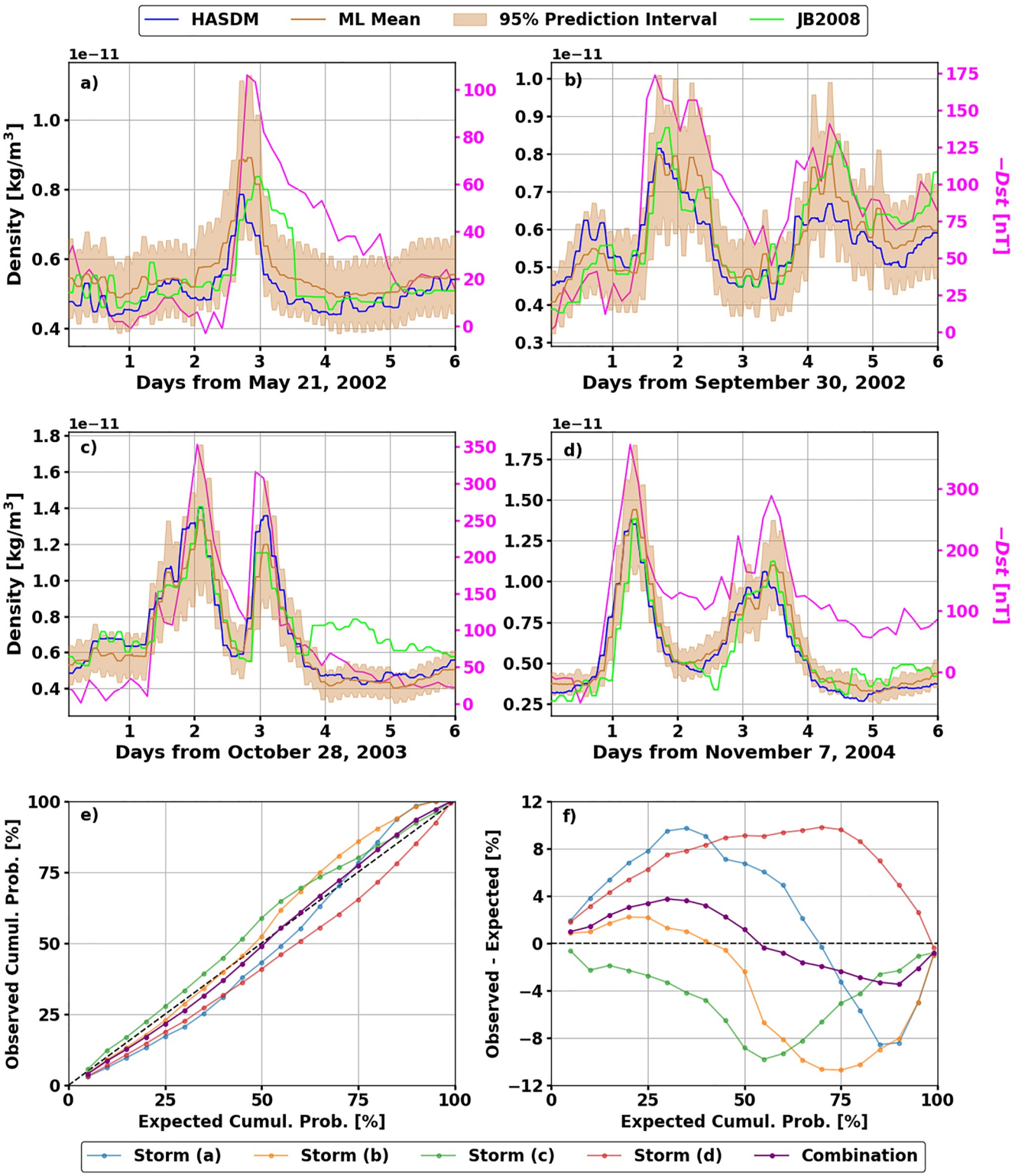

Figure 7. Panels (a), (b), (c), and (d) show High Accuracy Satellite Drag Model (HASDM), HASDM-ML mean, and JB2008 orbit-averaged density for CHAllenging Minisatellite Payload's orbit across various geomagnetic storms. The shaded region represents the 95% prediction interval for HASDM-ML and -storm-time disturbance index is shown on the right axis in each panel. Panel (e) shows the calibration curves corresponding to panels (a), (b), (c), and (d) along with the composite calibration curve (see bottom legend). Panel (f) shows the difference between the observed and expected cumulative probability for all the curves in panel (e).

meaningful unless they are well calibrated, and HASDM-ML is able to provide that. HASDM-ML was also more accurate than JB2008 relative to HASDM for all four storms and across the 20 space weather conditions considered (Table 6).

5. Limitations and Future Work

To further improve HASDM-ML, additional data can be introduced, particularly in the coming solar maximum. As seen in Section 3.1, periods of $ap > 50$ were the source of the highest overall error and the introduction of additional storms could reduce that error. Hierarchical modeling is another approach to potentially combat this limitation, as previously mentioned. We hypothesize that a nonlinear dimensionality reduction method could also improve performance, especially in modeling the highly nonlinear storm response (Turner et al., 2020). The developed model can be used in research (e.g., to study historical storms where HASDM outputs are unavailable) as well as in an operational setting. Another area of future work could be to develop a nonlinear dynamic ROM based on the SET HASDM density database. While the background model is static, the assimilative framework represents the dynamic thermosphere and a dynamic ROM could better model certain phenomena dealing with the dynamic evolution of the system. The dynamic ROM can also provide a framework for further data assimilation (Mehta & Linares, 2018).

Data Availability Statement

Requests can be submitted for full access to the Space environment technologies (SET) High Accuracy Satellite Drag Model density database at https://spacewx.com/hasdm/ and all reasonable requests for scientific research are accepted as explained in the rules of road document on the website. The historical space weather indices used in this study can be found at the following links: $F_{10.7}$: https://www.spaceweather.gc.ca/forecast-prevision/solar-solaire/solarflux/sx-en.php, planetary amplitude index (*ap*): https://doi.org/10.5880/Kp.0001, and storm-time disturbance index (*Dst*): https://doi.org/10.5880/Kp.0001. The remaining solar indices and proxies can be found at https://spacewx.com/jb2008/ in the SOLFSMY. TXT file. The primary source for nowcasts and forecasts of the solar drivers is provided by SET at the JB2008 link or on the Unified Data Library (UDL) at https://unifieddatalibrary.com/. Nowcasts and forecasts for *ap* are provided by the National Oceanic and Atmospheric Administration Space Weather Prediction Center and SET at the same links. The nowcasts and forecasts for *Dst* are provided by SET on UDL or at https://spacewx.com/new_sam_ops/. The forecasting models for these drivers have been benchmarked by Licata, Tobiska, and Mehta (2020). Free and one-time only registration is required to access the nowcasts and forecasts. CHAllenging Minisatellite Payload and Gravity Recovery and Climate Experiment position data were obtained from the measurements presented by Mehta et al. (2017) at http://tinyurl.com/densitysets.

Acknowledgments
This work was made possible by NASA Established Program to Stimulate Competitive Research, Grant #80NSSC19M0054. Space environment technologies and WVU gratefully acknowledge support from the NASA SBIR contract #80NSSC20C0292 for Machine learning Enabled Thermosphere Advanced by High Accuracy Satellite Drag Model (META-HASDM). We would like to thank Space Weather Canada for providing and maintaining solar radio emission data, GFZ Potsdam for supplying planetary amplitude index archives, and the World Data Center for Geomagnetism in Kyoto for providing storm-time disturbance index data. The authors would like to acknowledge NASA and DLR for their work in the CHAllenging Minisatellite Payload and Gravity Recovery and Climate Experiment missions along with GFZ Potsdam for managing the data. The authors would like to thank the anonymous reviewers for all of their time and effort. Their feedback allowed us to significantly improve the manuscript.

References

Abdelminaam, D. S., Ismail, F. H., Taha, M., Taha, A., Houssein, E. H., & Nabil, A. (2021). CoAID-DEEP: An optimized intelligent framework for automated detecting COVID-19 misleading information on twitter. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3058066

Alake, R. (2020). *Understanding and implementing dropout in TensorFlow and keras*. Retrieved from https://towardsdatascience.com/understanding-and-implementing-dropout-in-tensorflow-and-keras-a8a3a02c1bfa

Anderson, G. J., Gaffney, J. A., Spears, B. K., Bremer, P.-T., Anirudh, R., & Thiagarajan, J. J. (2020). *Meaningful uncertainties from deep neural network surrogates of large-scale numerical simulations*.

Bettadpur, S. (2012). Gravity Recovery and climate experiment: Product specification document, GRACE 327-720, CSR-GR-03-02, cent. for Space Res., The Univ. of Texas. Retrieved from https://podaac.jpl.nasa.gov/GRACE

Boniface, C., & Bruinsma, S. (2021). Uncertainty quantification of the DTM2020 thermosphere model. *Journal of Space Weather and Space Climate*, *11*, 53. https://doi.org/10.1051/swsc/2021034

Bowman, B., & Storz, M. (2003). High accuracy satellite drag model (HASDM) review. In *AIAA/AAS astrodynamics specialist conference*. AAS 03-625 Retrieved from https://sol.spacenvironment.net/~JB2008/pubs/JB2006_AAS_2003_625.pdf

Bowman, B., Tobiska, W. K., Marcos, F., Huang, C., Lin, C., & Burke, W. (2008). A new empirical thermospheric density model JB2008 using new solar and geomagnetic indices. In *AIAA/AAS astrodynamics specialist conference*. AIAA 2008-6438 Retrieved from https://arc.aiaa.org/doi/abs/10.2514/6.2008-6438

Bruinsma, S., & Biancale, R. (2003). Total densities derived from accelerometer data. *Journal of Spacecraft and Rockets*, *40*(2), 230–236. https://doi.org/10.2514/2.3937

Bruinsma, S., & Boniface, C. (2021). The operational and research DTM-2020 thermosphere models. *J. Space Weather Space Clim*, *11*, 47. https://doi.org/10.1051/swsc/2021032

Bruinsma, S., Boniface, C., Sutton, E. K., & Fedrizzi, M. (2021). Thermosphere modeling capabilities assessment: Geomagnetic storms. *J. Space Weather Space Clim*, *11*, 12. https://doi.org/10.1051/swsc/2021002

Bussy-Virat, C. D., Ridley, A. J., & Getchius, J. W. (2018). Effects of uncertainties in the atmospheric density on the probability of collision between space objects. *Space Weather*, *16*(5), 519–537. https://doi.org/10.1029/2017SW001705

Calabia, A., & Jin, S. (2016). New modes and mechanisms of thermospheric mass density variations from GRACE accelerometers. *Journal of Geophysical Research: Space Physics*, *121*(11), 11191–11212. https://doi.org/10.1002/2016JA022594

Camporeale, E., & Care, A. (2021). Accrue: Accurate and reliable uncertainty estimate in deterministic models. *International Journal for Uncertainty Quantification*, *11*(4), 81–94. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2021034623

Chandorkar, M., Camporeale, E., & Wing, S. (2017). Probabilistic forecasting of the disturbance storm time index: An autoregressive Gaussian process approach. *Space Weather*, *15*(8), 1004–1019. https://doi.org/10.1002/2017SW001627

Davison, A., Hinkley, D., Gill, R., Ripley, B., Ross, S., Stein, M., et al. (1997). *Bootstrap methods and their application*. In *Cambridge Series in Statistical and Probabilistic Mathematics* (pp. 243–244). Cambridge University Press.

Doornbos, E. (2012). *Producing density and crosswind data from satellite dynamics observations* (pp. 91–126). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-25129-0_4

Efron, B., & Tibshirani, R. (1993). *An Introduction to the bootstrap, monographs on statistics and applied probability* (pp. 338–339). Chapman & Hall.

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural architecture search: A survey. *Journal of Machine Learning Research*, *20*(55), 1–21. https://doi.org/10.1007/978-3-030-05318-5_11

Emmert, J. (2015). Thermospheric mass density: A review. *Advances in Space Research*, *56*(5), 773–824. https://doi.org/10.1016/j.asr.2015.05.038

Emmert, J., Warren, H., Segerman, A., Byers, J., & Picone, J. (2017). Propagation of atmospheric density errors to satellite orbits. *Advances in Space Research*, *59*(1), 147–165. https://doi.org/10.1016/j.asr.2016.07.036

Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning*. arXiv:1506.02142v6.

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118. https://doi.org/10.1175/MWR2904.1

Gondelach, D. J., & Linares, R. (2020). Real-time thermospheric density estimation via two-line element data assimilation. *Space Weather*, *18*(2), e2019SW002

Haegel, L., & Husa, S. (2020). Predicting the properties of black-hole merger remnants with deep neural networks. *Classical and Quantum Gravity*, *37*(13), 135005. https://doi.org/10.1088/1361-6382/ab905c

Holzinger, M. J., & Jah, M. K. (2018). Challenges and potential in space Domain awareness. *Journal of Guidance, Control, and Dynamics*, *41*(1), 15–18. https://doi.org/10.2514/1.G003483

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/10.1037/h0071325

Knipp, D., Tobiska, W. K., & Emery, B. (2004). Direct and indirect thermospheric heating sources for solar cycles 21-23. *Solar Physics*, *224*(1–2), 495–505. https://doi.org/10.1007/s11207-005-6393-4

Lei, J., Matsuo, T., Dou, X., Sutton, E., & Luan, X. (2012). Annual and semiannual variations of thermospheric density: EOF analysis of CHAMP and GRACE data. *Journal of Geophysical Research: Space Physics*, *117*(A1). https://doi.org/10.1029/2011JA017324

Licata, R., Mehta, P., & Tobiska, W. K. (2020). Impact of space weather driver forecast uncertainty on drag and orbit prediction. In *Proceedings of the 2020 AAS/AIAA astrodynamics specialist conference*. Retrieved from https://spacewx.net/wp-content/uploads/2020/08/ASC_201.pdf

Licata, R., Mehta, P., & Tobiska, W. K. (2021). Impact of driver and model uncertainty on drag and orbit prediction. In *Proceedings of the 31st AAS/AIAA space flight mechanics meeting*. Retrieved from https://www.researchgate.net/publication/349117733_Impact_of_Driver_and_Model_Uncertainty_on_Drag_and_Orbit_Prediction

Licata, R. J., Mehta, P. M., Tobiska, W. K., Bowman, B. R., & Pilinski, M. D. (2021). Qualitative and quantitative assessment of the SET HASDM database. *Space Weather*, *19*(8), e2021SW002. https://doi.org/10.1029/2021SW002798

Licata, R. J., Mehta, P. M., Weimer, D. R., & Tobiska, W. K. (2021). Improved neutral density predictions through machine learning enabled exospheric temperature model. *Journal Space Weather*, *19*(12), e2021SW002. https://doi.org/10.1029/2021SW002918

Licata, R. J., Tobiska, W. K., & Mehta, P. M. (2020). Benchmarking forecasting models for space weather drivers. *Space Weather*, *18*(10), e2020SW002. https://doi.org/10.1029/2020SW002496

Liu, H., Lühr, H., Henize, V., & Köhler, W. (2005). Global distribution of the thermospheric total mass density derived from CHAMP. *Journal of Geophysical Research: Space Physics*, *110*(A4). https://doi.org/10.1029/2004JA010741

Luhr, H., Grunwaldt, L., & Forste, C. (2002). *CHAMP reference systems, transformations and standards, tech. Rep*. CH-GFZ-RS-002, GFZ-Potsdam Retrieved from http://www-app2.gfz-potsdam.de/pb1/op/champ/more/docs_CHAMP.html

March, G., Doornbos, E., & Visser, P. (2019). High-fidelity geometry models for improving the consistency of CHAMP, GRACE, GOCE and Swarm thermospheric density data sets. *Advances in Space Research*, *63*(1), 213–238. https://doi.org/10.1016/j.asr.2018.07.009

March, G., van den IJssel, J., Siemes, C., Visser, P. N. A. M., Doornbos, E. N., & Pilinski, M. (2021). Gas-surface interactions modelling influence on satellite aerodynamics and thermosphere mass density. *Journal of Space Weather and Space Climate*, *11*, 54. https://doi.org/10.1051/swsc/2021035

Marcos, F., Kendra, M., Griffin, J., Bass, J., Larson, D., & Liu, J. J. (1998). Precision low Earth orbit determination using atmospheric density calibration. *Journal of the Astronautical Sciences*, *46*(4), 395–409. https://doi.org/10.1007/BF03546389

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096. Retrieved from http://www.jstor.org/stable/2629907

Matsuo, T., & Forbes, J. M. (2010). Principal modes of thermospheric density variability: Empirical orthogonal function analysis of CHAMP 2001–2008 data. *Journal of Geophysical Research: Space Physics*, *115*(A7). https://doi.org/10.1029/2009JA015109

Mehta, P. M., & Linares, R. (2017). A methodology for reduced order modeling and calibration of the upper atmosphere. *Space Weather*, *15*(10), 1270–1287. https://doi.org/10.1002/2017SW001642

Mehta, P. M., & Linares, R. (2018). A new transformative framework for data assimilation and calibration of physical ionosphere-thermosphere models. *Space Weather*, *16*(8), 1086–1100. https://doi.org/10.1029/2018SW001875

Mehta, P. M., & Linares, R. (2020). Real-time thermospheric density estimation from satellite position measurements. *Journal of Guidance, Control, and Dynamics*, *43*(9), 1656–1670. https://doi.org/10.2514/1.G004793

Mehta, P. M., Linares, R., & Sutton, E. K. (2018). A quasi-physical dynamic reduced order model for thermospheric mass density via hermitian space-dynamic mode decomposition. *Space Weather*, *16*(5), 569–588. https://doi.org/10.1029/2018SW001840

Mehta, P. M., Walker, A. C., Sutton, E. K., & Godinez, H. C. (2017). New density estimates derived using accelerometers on board the CHAMP and GRACE satellites. *Space Weather*, *15*(4), 558–576. https://doi.org/10.1002/2016SW001562

Muelhaupt, T. J., Sorge, M. E., Morin, J., & Wilson, R. S. (2019). Space traffic management in the new space era. *Journal of Space Safety Engineering*, *6*(2), 80–87. https://doi.org/10.1016/j.jsse.2019.05.007

Nazarenko, A., Cefola, P., & Yurasov, V. (1998). Estimating atmospheric density variations to improve LEO orbit prediction accuracy. In *AIAA/AAS space flight mechanics meeting*. AAS 98-190 Retrieved from http://www.space-flight.org/AAS_meetings/1998_winter/abstracts/98-190.html

Oliveira, D. M., Zesta, E., Schuck, P. W., & Sutton, E. K. (2017). Thermosphere global time response to geomagnetic storms caused by coronal mass ejections. *Journal of Geophysical Research: Space Physics*, *122*(12). https://doi.org/10.1002/2017JA024006

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). *Keras tuner*. Retrieved from https://github.com/keras-team/keras-tuner

Palmroth, M., Grandin, M., Sarris, T., Doornbos, E., Tourgaidis, S., Aikio, A., et al. (2021). Lower-thermosphere–ionosphere (LTI) quantities: Current status of measuring techniques and models. *Annales Geophysicae*, *39*(1), 189–237. https://doi.org/10.5194/angeo-39-189-2021

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pevec, D., & Kononenko, I. (2014). Prediction intervals in supervised learning for model evaluation and discrimination. *Applied Intelligence*, *42*(4), 790–804. https://doi.org/10.1007/s10489-014-0632-z

Picone, J. M., Hedin, A. E., Drob, D. P., & Aikin, A. C. (2002). NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *Journal of Geophysical Research: Space Physics*, *107*. https://doi.org/10.1029/2002JA009430

Radtke, J., Kebschull, C., & Stoll, E. (2017). Interactions of the space debris environment with mega constellations–Using the example of the OneWeb constellation. *Acta Astronautica*, *131*, 55–68. https://doi.org/10.1016/j.actaastro.2016.11.021

Rasmussen, C. E. (2004). *Gaussian processes in machine learning* (pp. 63–71). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_4

Rogachev, A. F., & Melikhova, E. V. (2020). Automation of the process of selecting hyperparameters for artificial neural networks for processing retrospective text information. *IOP Conference Series: Earth and Environmental Science*, *577*(1), 012012. https://doi.org/10.1088/1755-1315/577/1/012012

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. Retrieved from http://jmlr.org/papers/v15/srivastava14a.html

Storz, M. F., Bowman, B. R., Branson, M. J. I., Casali, S. J., & Tobiska, W. K. (2005). High accuracy satellite drag model (hasdm). *Advances in Space Research*, *36*(12), 2497–2505. https://doi.org/10.1016/j.asr.2004.02.020

Sutton, E. K. (2008). *Effects of solar disturbances on the thermosphere densities and winds from CHAMP and GRACE satellite accelerometer data*. Ph.D. thesis, University of Colorado at Boulder.

Tobiska, W. K., Bouwer, S. D., & Bowman, B. R. (2008). The development of new solar indices for use in thermospheric density modeling. *Journal of Atmospheric and Solar-Terrestrial Physics*, *70*(5), 803–819. https://doi.org/10.1016/j.jastp.2007.11.001

Tobiska, W. K., Bowman, B. R., Bouwer, D., Cruz, A., Wahl, K., Pilinski, M., et al. (2021). *The SET HASDM density database*. Space Weather. e2020SW002682. https://doi.org/10.1029/2020SW002682

Turner, H., Zhang, M., Gondelach, D., & Linares, R. (2020). Machine learning algorithms for improved thermospheric density modeling. In F. Darema, E. Blasch, S. Ravela, & A. Aved (Eds.), *Dynamic data driven applications systems* (pp. 143–151). Springer International Publishing.

Weimer, D. R., Mehta, P. M., Tobiska, W. K., Doornbos, E., Mlynczak, M. G., Drob, D. P., & Emmert, J. T. (2020). Improving neutral density predictions using exospheric temperatures calculated on a geodesic, polyhedral grid. *Space Weather*, *18*(1), e2019SW002. https://doi.org/10.1029/2019SW002355