2022

Probabilistic Space Weather Modeling and Forecasting for the Challenge of Orbital Drag in Space Traffic Management

Richard J. Licata III
West Virginia University, rjlicata@mix.wvu.edu

Probabilistic Space Weather Modeling and Forecasting for the Challenge of Orbital Drag in Space Traffic Management

Richard J. Licata

DISSERTATION submitted to the

Benjamin M. Statler College of Engineering and Mineral Resources

at West Virginia University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY in **Aerospace Engineering**

Piyush M. Mehta, Ph.D., Chair

Jason Gross, Ph.D.

Yu Gu, Ph.D.

Snehalata Huzurbazar, Ph.D.

Nasser Nasrabadi, Ph.D.

Natalia Schmid, Ph.D.

Department of Mechanical and Aerospace Engineering

Morgantown, West Virginia

2022

Keywords: satellite drag, machine learning, space weather,

thermosphere, collision avoidance, uncertainty

ABSTRACT

Probabilistic Space Weather Modeling and Forecasting for the Challenge of Orbital Drag in Space Traffic Management

Richard J. Licata

In the modern space age, private companies are crowding the already-congested low Earth orbit (LEO) regime with small satellite mega constellations. With over 25,000 objects larger than 10 cm already in LEO, this rapid expansion is forcing us towards the enterprise on Space Traffic Management (STM). STM is an operational effort that focuses on conjunction assessment and collision avoidance between objects. While the equations of motion for objects in orbit are well-known, there are many uncertain parameters that result in the uncertainty of an object's future position. The force that the atmosphere exerts on satellite – known as drag – is the largest source of uncertainty in LEO. This is largely due to the difficulty in predicting mass density in the thermosphere – the neutral region in Earth's upper atmosphere. Presently, most thermosphere models are deterministic and the treatment of uncertainty in density is highly simplified or nonexistent in operations.

In this work, four probabilistic thermospheric mass density models are developed using machine learning (ML) to enable the investigation of the impact of model uncertainty on satellite position for the first time. Of these four models, two (HASDM-ML and TIE-GCM ROPE) are reduced order models based on outputs from existing thermosphere models while the other two (CHAMP-ML and MSIS-UQ) are based on in-situ thermosphere measurements. The data and model development are described, and the models' capabilities, including the robustness of their uncertainty quantification (UQ) capabilities, are thoroughly assessed.

Existing thermosphere models, and the ones developed here, use different space weather drivers to estimate density. In a forecasting environment, there are algorithms and models that forecast the

drivers for a given period in order for a density model to make a forecast. The driver forecast models used by the United States Space Force for the HASDM system are assessed to benchmark our current capabilities. Using the error statistics for each driver, we can perturb the deterministic forecasts. This provides an avenue to use the ML thermosphere models to study the effect of driver uncertainty on satellite position, in addition to model uncertainty, for any period with available driver forecasts. Seven periods are considered with diverse space weather conditions to study the isolated effects of the two density uncertainty sources on a 72-hour satellite orbit. This provides insight into the relative importance of density uncertainty on satellite position for various space weather scenarios. This study also functions as a motivation to reconsider our current methods for STM in order to improve our capabilities and prevent future satellite collisions with increased confidence.

Table of Contents

List of Figures

List of Tables

List of Publications

Portions of this dissertation's content originally appeared in some of the following journal and conference publications.

Peer-Reviewed

1 **Licata, R. J.**, Tobiska, W. K., and Mehta, P. M. (2020) Benchmarking Forecasting Models for Space Weather Drivers. *Space Weather*, *18*, e2020SW002496. https://doi.org/10.1029/2020SW002496

2 Tobiska, W. K., Bowman, B. R., Bouwer, S. D., Cruz, A., Wahl, K., Pilinski, M. D., Mehta, P. M., and **Licata, R. J.** (2021) The SET HASDM Density Database. *Space Weather*, *19*, e2020SW002682. https://doi.org/10.1029/2020SW002682

3 **Licata, R. J.**, Mehta, P. M., Tobiska, W. K., Bowman, B. R., and Pilinski, M. D. (2021) Qualitative and Qualitative Assessment of the SET HASDM Database, *Space Weather*, *19*, e2021SW002798, https://doi.org/10.1029/2021SW002798.

4 **Licata, R. J.**, Mehta, P. M., Weimer, D. R., and Tobiska, W. K. (2021) Improved Neutral Density Predictions through Machine Learning Enabled Exospheric Temperature Model, *Space Weather*, *19*, e2021SW002918, https://doi.org/10.1029/2021SW002918.

5 Oliveira, D. M., Zesta, E., Mehta, P. M., **Licata, R. J.**, Pilinski, M. D., Tobiska, W. K., and Hayakawa, H. (2021) The Current and Future Directions of Modeling Thermosphere Density Enhancements during Extreme Magnetic Storms. *Frontiers in Astronomy and Space Science*, *8*. https://doi.org/10.3389/fspas.2021.764144

6 Weimer, D. R., Tobiska, W. K., Mehta, P. M., **Licata, R. J.**, and Drob, D. P. (2021) Comparison of a Neutral Density Model with the SET HASDM Density Database, Space Weather, 19, e2021SW002888. https://doi.org/10.1029/2021SW002888.

7 **Licata, R. J.**, Mehta, P. M., Tobiska, W. K., and Hurzurbazar, S. (2021) Machine-Learned HASDM Model with Uncertainty Quantification, *Space Weather*, 20, e2021SW002915. https://doi.org/10.1029/2021SW002915.

8 **Licata, R. J.** and Mehta, P. M. (2022) Uncertainty Quantification Techniques for Space Weather Modeling: Thermospheric Density Application, *Scientific Reports*, 12, 7256. https://doi.org/10.1038/s41598-022-11049-3.

9 **Licata, R. J.**, Mehta. P. M. , Weimer, D. R., Drob, D. P., Tobiska, W. K., and Yoshii, J. (2022) Science through Machine Learning: Quantification of Poststorm Thermospheric Cooling, *Space Weather*, 20, e2022SW003189. https://doi.org/10.1029/2022SW003189.

10 **Licata, R. J.**, Mehta. P. M. , Weimer, D. R., Tobiska, W. K., and Yoshii, J. (2022) MSIS-UQ: Calibrated and Enhanced NRLMSIS 2.0 Model with Uncertainty Quantification, *Space Weather*, 20, e2022SW003267. https://doi.org/10.1029/2022SW003267.

11 Paul, S. N., **Licata, R. J.**, and Mehta. P. M. (2022) Advanced ensemble modeling method for space object state prediction accounting for uncertainty in atmospheric density, *Under Review, Advances in Space Research*. https://doi.org/10.48550/arXiv.2210.16992.

12 Paul, S. N., Sheridan, P. L., **Licata, R. J.**, and Mehta. P. M. (2022) Stochastic modeling of physical drag coefficient - its impact on orbit prediction and space traffic management, *Under Review, Advances in Space Research*. https://doi.org/10.48550/arXiv.2210.08364.

13 **Licata, R. J.** and Mehta, P. M. (2022) Reduced Order Probabilistic Emulation for Physics-Based Thermosphere Models, *Under Review, Space Weather*. https://doi.org/10.48550/arXiv.2211.04392.

Conference Proceedings

Conference Papers

14 **Licata, R. J.**, Mehta, P. M., and Kay, C. (2019) Data-Driven Framework for Space Weather Modeling with Uncertainty Treatment Towards Space Situational Awareness and Space Traffic Management, in *Proceedings of the 2019 AAS/AIAA Astrodynamics Specialist Conference*.

15 **Licata, R. J.**, Mehta, P. M., and 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*.

16 **Licata, R. J.**, Mehta, P. M., and 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*.

Conference Posters and Oral Presentations

17 **Licata, R. J.** and Mehta, P. M. (2019) Physics-informed Machine Learning for Probabilistic Space Weather Modeling and Forecasting: Thermosphere and Satellite Drag, in *Proceedings of AGU Fall Meeting 2019*.

18 Mehta, P. M., **Licata, R. J.**, Welling D. T., Cash, M., and Morley, S. K. (2019) Physics-informed Machine Learning fo Probabilistic Space Weather Modeling and Forecasting: dB/dt and Geomagnetically Induced Currents, in *Proceedings of AGU Fall Meeting 2019*.

19 **Licata, R. J.** and Mehta, P. M. (2020) Physics-informed Machine Learning with Autoencoders and LSTM for Probabilistic Space Weather Modeling and Forecasting, in *Proceedings of the 17th Conference on Space Weather – 100th AMS Annual Meeting*.

20 **Licata, R. J.**, Mehta, P. M., and W. Kent Tobiska (2020) Data-Driven HASDM Density Model using Machine Learning, in *Proceedings of AGU Fall Meeting 2020*.

21 **Licata, R. J.**, Mehta, P. M., Cruz, A. A., Kirk, M., and Thompson, B. J. (2021) Using Deep Learning for Onboard Data Compression on Deep Space Missions, in *Proceedings of the 18th Conference on Space Weather – 101st AMS Annual Meeting*.

22 **Licata, R. J.** and Mehta, P. M. (2021) Physics-informed Bayesian Deep Learning for Space Weather Science and Operations, in *Proceedings of the 18th Conference on Space Weather – 101st AMS Annual Meeting*.

23 Tobiska, W. K., Bowman, B. R., Pilinski, M. D., Mehta, P. M., and **Licata, R. J.** (2021) The Machine Learning Enabled Thermosphere Advanced by the High Accuracy Satellite Drag Model (META-HASDM), in *Proceedings of The Advanced Maui Optical and Space Surveillance Technologies (AMOS) Conference 2021*.

24 **Licata, R. J.**, Mehta, P. M., Weimer, D. R., and Tobiska, W. K. (2021) Improved Neutral Density Predictions through Machine Learning Enabled Exospheric Temperature Model, in *Proceedings of AGU Fall Meeting 2021*.

25 **Licata, R. J.**, Mehta, P. M., Tobiska, W. K., and Hurzurbazar, S. (2022) Machine-Learned HASDM Model with Uncertainty Quantification, in *Proceedings of the 19th Conference on Space Weather – 102nd AMS Annual Meeting*.

Acknowledgments

I want to start by thanking my advisor, Dr. Piyush Mehta, for taking a chance on me my senior year, giving me an offer to be a GRA and pushing me to pursue a Ph.D, a dream I did not even know I had. You have pushed me to be the best researcher I can be and set me up for my future career. I appreciate your focus early on to improve my writing skills and for immediately throwing me into the deep end with my first conference paper. I have enjoyed our discussions and debates about space weather, machine learning, and life. I cannot thank you enough for the opportunity you provided me.

Thank you to all the members of my committee: Dr. Jason Gross, Dr. Yu Gu, Dr. Snehalata Huzurbazar, Dr. Nasser Nasrabadi, and Dr. Natalia Schmid. I truly appreciate your time spent reviewing this dissertation, providing feedback, and answering any questions I have had. I also want to thank Dr. Andrew Rhodes for always being available to discuss the Ph.D. process.

I would like to thank a few others who have provided guidance in my time as a graduate student. Thank you Dr. Kent Tobiska for including me on various research projects over the years and giving insight for my research, particular related to the HASDM project. Thank you Dr. Daniel Weimer for your guidance and inclusion on the EXTEMPLAR project for the past few years. Thank you to other project members such as Dr. Marcin Pilinski, Dr. Douglas Drob, and Dr. Jean Yoshii who contributed to discussions that impacted my graduate research. I would also like to thank the other members of Dr. Mehta's research group who were always available for calls, would enjoy spirited debates on research topics, and fostered a positive environment.

Thank you to my parents, sisters, grandparents, and in-laws for providing crucial support through my graduate studies. Continued phone calls and check-ins over the past few years have helped keep me grounded through stressful times. Of course, I must thank my Poppop specifically for keeping things light and offering brilliant ideas such as "traffic warning signs" in space.

Last but definitely not least, I want to thank my beautiful wife, Nora. Meeting you was a blessing, and although our future was uncertain with your graduation, we were able to make it work. I am forever thankful that you moved back to Morgantown while I pursued this degree and were by my side through the highs and lows of this process. I could not have made it to this point without your endless love, support, and patience.

Nomenclature

Symbol	Definition	Units
A	cross-sectional area	m^2
a_{drag}	drag acceleration	m/s^2
ap	geomagnetic 3-hourly planetary equivalent amplitude index	$2nT$
Ap	geomagnetic daily planetary amplitude index	$2nT$
B	ballistic coefficient	m^2/kg
B_Z	north-south component of the IMF	nT
C_D	drag coefficient	–
CO_2	carbon dioxide	–
Dst	disturbance storm time index	nT
F_{10}	solar radio flux measured at 10.7 cm wavelength	sfu
H	atomic hydrogen	–
He	helium	–
$\mathbb{I}$	indicator function	–
k	number of Monte Carlo samples	–
Kp	global logarithmic geomagnetic activity index	–
$\mathcal{L}$	training loss	–

m	mass	kg
M_{10}	proxy for FUV photospheric 160 nm Schumann-Runge Continuum emissions	sfu
n	number of samples	–
N	atomic nitrogen	–
N_2	molecular nitrogen	–
n_{inp}	number of inputs	–
n_{ip}	number of initial points	–
n_{LS}	number of lag steps	–
NO	nitric oxide	–
n_{out}	number of outputs	–
O	atomic oxygen	–
O_2	molecular oxygen	–
p	probability	–
r	choice order of truncation for PCA	–
R^2	coefficient of determination	–
S_{10}	index for integrated 26-34 nm bandpass solar chromospheric EUV emission	sfu
sfu	solar flux units	–
S_N	Poynting flux in the Northern hemisphere	GW
S_S	Poynting flux in the Southern hemisphere	GW
SYM-H	longitudinally symmetric component of magnetic field disturbances	*nT*

ΔT	exospheric temperature perturbation	K
T_∞	exospheric temperature	K
T_{120}	temperature at 120 km	K
T'_{120}	vertical gradient at 120 km	K/m
v_{rel}	relative velocity	m/s
w_t	model weights at time, t	–
Y_{I0}	hybrid solar activity index	sfu

Greek Letters	**Definition**	**Units**
α	PCA coefficient	–
β	first-order Gauss-Markov process parameter	–
θ	input/output variables	–
$\widetilde{\theta}$	scaled input/output variables	–
η	learning rate	–
κ	bias factor	–
μ	mean	–
ρ	thermospheric mass density	kg/m^3
σ	standard deviation	–
σ^2	variance	–
τ	half-life	min

ϕ	input to activation function	–

Abbreviations	**Definition**
ACE	Advanced Composition Explorer
ANN	artificial neural network
cdf	cumulative distribution function
CES	calibration error score
CHAMP	CHAllenging Minisatellite Payload
CI	confidence interval
CME	coronal mass ejection
CNN	convolutional neural network
CRPS	continuous ranked probability score
CTIPe	Coupled Thermosphere Ionosphere Plasmasphere Electrodynamic model
DCA	Dynamic Calibration of the Atmosphere
DCP	dynamic consider parameter
DMD	Dynamic Mode Decomposition
DMDc	Dynamic Mode Decomposition with control
DoD	Department of Defense
doy	day of year
DSCOVR	Deep Space Climate Observatory

DTM	Drag Temperature Model
EBM	error bound for the population mean
ELU	exponential linear unit
erf	Gauss error function
EUV	extreme ultraviolet
EXTEMPLAR	EXospheric TEMperatures on a PoLyhedrAl gRid
FUV	far ultraviolet
GITM	Global Ionosphere-Thermosphere Model
GOCE	Gravity Field and Steady-State Ocean Circulation Explorer
GP	Gaussian process
GPS	Global Positioning System
GPU	graphical processing unit
GRACE	Gravity Recovery and Climate Explorer
GRACE-FO	GRACE-Follow On
HASDM	High Accuracy Satellite Drag Model
HSS	high-speed stream
IC	image credit
ISO	International Organization for Standardization
IMF	interplanetary magnetic field
JB2008	Jacchia-Bowman 2008 Empirical Thermospheric Density Model

LEO	low-Earth orbit
LST	local solar time
LSTM	long short-term memory neural network
MAE	mean absolute error
MC	Monte Carlo
MIT	magnetosphere – ionosphere – thermosphere
ML	machine learning
MSE	mean square error
MSIS	Mass Spectrometer and Incoherent Scatter radar
NASA	National Aeronautics and Space Administration
NLPD	negative logarithm of predictive density
NOAA	National Oceanic and Atmospheric Administration
NRL	Naval Research Laboratory
PCA	principal component analysis
pdf	probability density function
PoC	Probability of Collision
PI	prediction interval
ReLU	rectified linear unit
RNN	recurrent neural network
ROPE	Reduced Order Probabilistic Emulator

ROM	reduced order model
RSO	resident space object
SABER	Sounding of the Atmosphere using Broadband Emission Radiometry
SDA	space domain awareness
SELU	scaled exponential linear unit
SET	Space Environment Technologies
SSA	space situational awareness
SSB	segmented solution for the ballistic coefficient
STM	space traffic management
SVD	singular value decomposition
SWPC	Space Weather Prediction Center
TAD	traveling atmospheric disturbance
TID	traveling ionospheric disturbance
TGCM	Thermosphere General Circulation Model
TIE-GCM	Thermosphere-Ionosphere-Electrodynamics General Circulation Model
USAF	United States Air Force
USSC	United Stated Space Command
USSF	United Stated Space Force
UT	universal time
UQ	uncertainty quantification

Chapter 1. Motivation

Space situational awareness (SSA) has long involved the process of detecting, tracking, and identifying all artificial objects in Earth orbit, an activity also known as catalog maintenance. While the long-term characterization of the orbital debris environment is of primary importance for space sustainability, there is also a need to assess the immediate risk on the scale of hours to days. This need for a more active and real-time knowledge of the space environment driven by the ever increasing congestion and contest in the domain has shifted the focus to space domain awareness (SDA) putting stress on the ability to accurately predict the state of resident space objects (RSOs). Accurate modeling of orbital perturbations is crucial to achieving the primary goals of SDA including remote sensing applications: RSO characterization, tracking, and prediction.

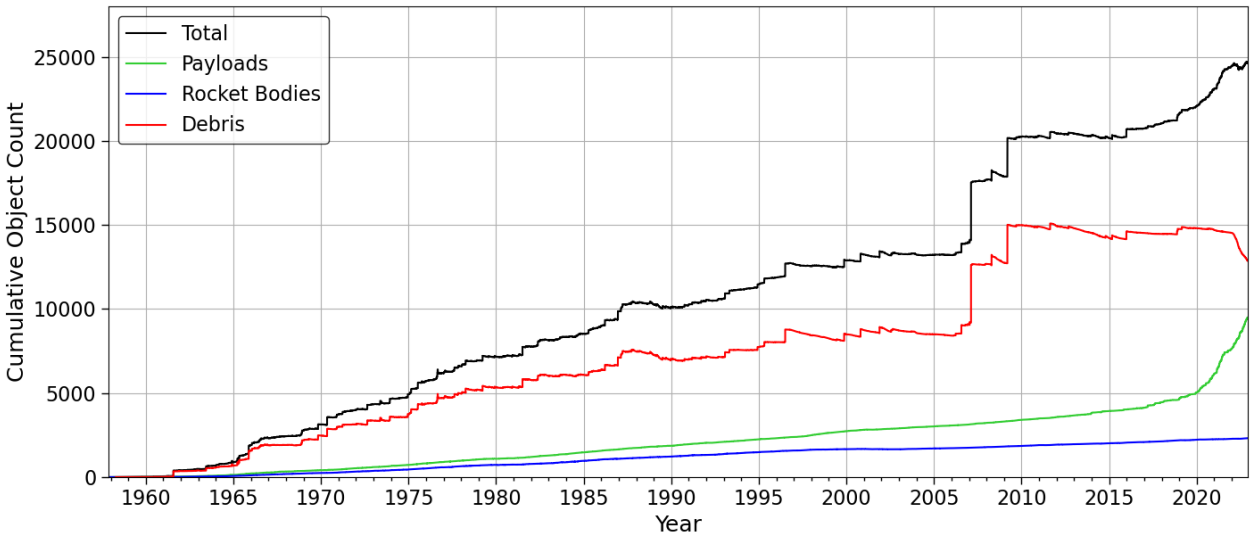

Figure 1.1: Population of satellite payload, rocket bodies, and debris since the launch of Sputnik in 1957. This data was retrieved from https://www.space-track.org/ [1].

The proliferation of low Earth orbit (LEO) with replenishable small satellite megaconstellations has shifted the focus from SSA and SDA to space traffic management (STM). The recent

sharp rise in payloads seen in Figure 1.1 is a result of the increasing prevalence of these megaconstellations. In addition to catalog maintenance under SSA, STM has put focus on enhanced and concerted space operations that includes conjunction assessment and collision avoidance [2]. The United States Space Command (USSC), now under the United State Space Force (USSF), maintains the most comprehensive public catalog of space objects – including debris – and is currently tasked with operations for the Department of Defense (DoD) and NASA. Presently, it provides collision warning messages with Probability of Collision (PoC) and best estimates of state parameters and covariance to most, if not all, operators. The operators either use this information at face value or deploy their own conjunction tools and gather additional actionable intelligence for making decisions since maneuvers are expensive (e.g. personnel cost, scientific data or commercial service outage, fuel costs) [3]. Additionally, the USSC and operators conduct their analyses using different space weather models making the process inconsistent when combining information. The continued LEO proliferation will overwhelm the current operational system and make decisions challenging as the operators are likely to receive multiple collision warning messages a day. For example, in the 6 month period between December 2021 and May 2022, SpaceX Starlink performed nearly 7,000 collision avoidance maneuvers [4].

1.1 The Challenge of Orbital Drag

In order to improve our ability to make confident decisions about potential collisions, we examine our current approaches to drag modeling. The primary sources of error, or uncertainty, in the drag acceleration (a_{drag}) model described below are the thermospheric mass density, ρ, and the drag coefficient, C_D.

$$a_{\text{drag}} = -\frac{1}{2}\rho B v_{\text{rel}}^2 \qquad \text{where} \qquad B = \frac{C_D A}{m} \tag{1.1}$$

Cross-sectional area, A, and satellite mass, m, are typically well known but can also be uncertain depending on the object. As a result, they are commonly lumped into a single uncertain ballistic coefficient parameter, B, to simplify the modeling.The final parameter is the velocity of the orbiting

object with respect of the co-rotating atmosphere, v_{rel}, which is also generally well known but can induce errors in the presence of strong neutral winds [5].

The effect of the thermosphere on satellite drag is a well-known problem in the space weather research and space operations communities. A number of empirical and physics-based thermosphere models have been developed (see Section 2.3.1), and significant efforts over the last two decades have reduced the mean global error of empirical models to sub-10% level during peak levels of the solar cycle. However, the errors during magnetically active conditions can be upwards of 25% [6]. Physics-based models can model the storm conditions with higher fidelity and potentially more accuracy but are computationally expensive and can be biased. Figure 1.2 illustrates how the Sun-Earth system are coupled and how modeling errors result in orbit prediction errors.

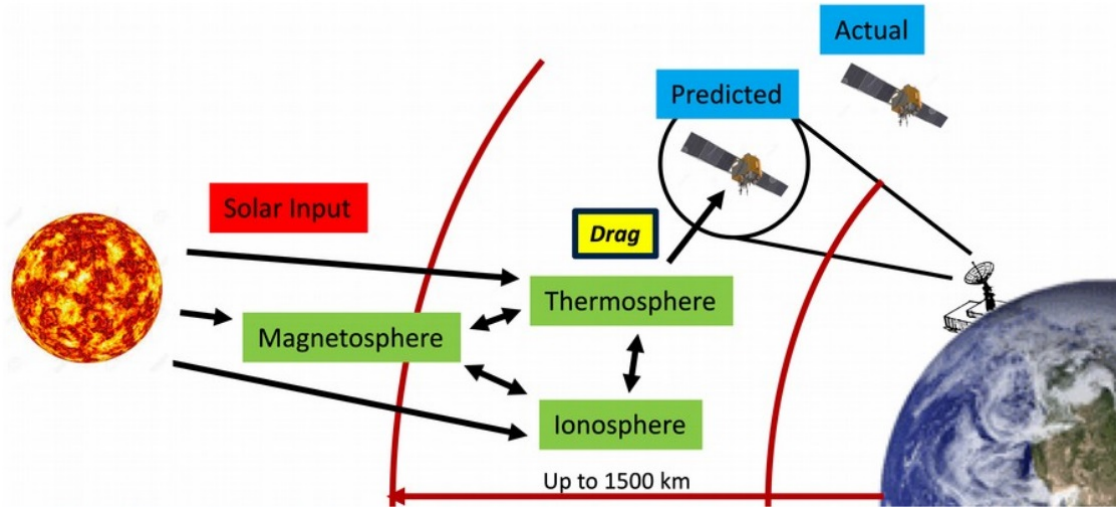

Figure 1.2: Coupling between space weather and thermosphere/Drag and its impact on orbit prediction.

While historical performance is a necessary indicator of density model forecast performance, it is not sufficient since forecasting is complicated by the requirement of consistency between density and drag/ballistic coefficient, as defined in Equation 1.1. Additionally, and critically for operations, quantification of uncertainty in thermosphere forecasts remains a major challenge. This is largely

due to inaccurate specification of processes (referred to here as model error or uncertainty) and driver forecasts (referred to here as driver uncertainty). The USSC and other service providers or operators typically perform a 3-day forecast to identify possible conjunction events. For the 3-day forecasts of LEO orbits, the position covariance or uncertainty has been assumed to be dominated by uncertainty in thermosphere model drivers.

The final piece of the puzzle is how to reliably and realistically map the effect of thermosphere model and driver uncertainty (and other uncertain parameters, including the drag coefficient) to the state and covariance of an orbiting object such that decisions can be made with confidence. The USSC currently uses a one-dimensional dynamic consider parameter (DCP) approach in prediction to map the time-invariant uncertainty at epoch to the forecast time using a state transition matrix. Additionally, uncertainty in the ballistic coefficient is accounted for through a frontal area factor, details of which are not public, and also mapped to the forecast using the DCP approach. Since there is no coupling between orbit determination and prediction, the overall covariance is calculated simply as a superposition of the mapped DCP(s) and propagated state covariance. As a result, the accuracy and realism of the resulting covariance can be significantly improved.

To summarize some of the major challenges associated with STM in the current environment, there are inconsistencies in operational methods and assumptions made that can result in inaccurate uncertainties for different objects. On the operations front, there is a lack of consistency when it comes to the models and tools used in decision making. This can cause operators from different agencies/companies to come to different conclusions for the same conjunction event. When it comes to modeling, the satellite drag coefficient and thermospheric density often carry high errors, but their respective uncertainties are often overlooked or oversimplified. This dissertation will focus solely on the problem of improving thermospheric density modeling, introducing model uncertainty, and studying the relative impact of model and driver uncertainty on satellite state forecasts.

1.2 Contributions

1.2.1 Probabilistic Thermospheric Mass Density Models

In Chapter 4, machine-learned thermosphere models are developed with the goals of computational efficiency, accessibility, and uncertainty quantification. CHAMP-ML is based on direct density estimates from the CHAMP satellite meaning CHAMP-ML learns a relationship between drivers and density without any underlying basis functions. This model is developed as a framework for a more universally applicable model based entirely on satellite data. HASDM-ML is unique, because the model it is based on is inaccessible to the public. HASDM-ML has less than 10% mean error relative to its original model and provides robust and reliable uncertainty estimates. MSIS-UQ is an exospheric temperature while all others developed in this work predict density. However, this is a key parameter in the MSIS formulation of mass density, so the improved accuracy in T_∞ predictions from MSIS-UQ allow for improved density prediction accuracy the standalone MSIS model. TIE-GCM ROPE is an ensemble-based reduced order probabilistic emulator for a computationally expensive physics-based model, TIE-GCM. The creation of this model allows for the incorporation of physical system dynamics into a ML model with uncertainty estimation capabilities. This work marks the first time probabilistic thermosphere models are developed with demonstrated reliability of uncertainty estimation capabilities

1.2.2 Extracting Science through Machine Learning

ML models are universal function approximators and – if used correctly – can summarize the information content of observational datasets in a functional form for scientific and engineering applications. A benefit to ML over parametric models is that there are no a priori assumptions about particular basis functions which can potentially limit the phenomena that can be modeled. The models developed in this work are used to study the presence of post-storm thermospheric overcooling in the middle-thermosphere in Chapter 5. This can be difficult to study through observations from the variability of other parameters, but ML can help us clearly identify its presence. This approach can be adopted in the future to answer other outstanding questions in the community.

1.2.3 Benchmarking Operational Space Weather Driver Forecasting Capabilities

Although much of the focus in density and drag modeling focuses on improving density models, our ability to accurately forecast their drivers is often overlooked. Even if we get to a point that we could predict thermospheric mass density with perfect accuracy given a set of model drivers, our inability to accurately forecast these drivers limits the reliability of forecasted density and therefore drag. In Chapter 6, the current forecasting models for six space weather drivers used by JB2008 and HASDM are benchmarked in order for others in the community to compare newly developed driver forecast models. These error statistics can also be used to perturb deterministic driver forecasts (see Chapter 7) although more robust methods are likely required (see Section 8.1).

1.2.4 Quantification of Driver and Model Uncertainty on Orbital State

Collision probability is often computed without the uncertainty in atmospheric drag taken into account. In this work, we investigate the isolated effects of driver and model uncertainty on a satellite state to highlight the importance of drag uncertainty in *PoC* calculations. The framework proposed and models developed provide a unique opportunity to show the relative importance of both uncertainties for the first time and why they both need to be accounted for in operations. Chapter 7 shows that **both** uncertainties can cause positions to be uncertain on the order of kilometers after only 72 hours, assuming no initial state uncertainty.

Chapter 2. Thermosphere and Space Weather

This chapter aims to provide the necessary background for the thermosphere and how it is impacted by the Sun and space weather. This information is crucial to understanding why thermospheric mass density is such an important and uncertain parameter in collision risk assessment.

2.1 Thermospheric Neutral Mass Density

The thermosphere is the neutral portion of Earth's upper atmosphere, and it ranges from $\sim$90 km to 500-1000 km depending on space weather conditions. Variations within it are primarily related to temperature responses. The primary heating source for the thermosphere is the absorption of solar extreme ultraviolet (EUV) and far ultraviolet (FUV) irradiance [7]. As the amount of solar irradiance changes, the thermosphere will either expand or contract as it is heated or cooled, respectively [8]. This effect essentially provides the baseline average thermospheric mass density [9]. The amount of solar irradiance, as well as its relative importance, fluctuates across the solar cycle, an eleven-year period where the magnetic activity of the Sun cycles in strength. The two main phases of the cycle are when the Sun is most active (solar maximum) and least active (solar minimum).

During solar minimum, an important contributing factor to density variations in the thermosphere is the continuing change to its composition [10]. The major constituents of the neutral thermosphere are atomic nitrogen (N), molecular nitrogen (N_2), atomic oxygen (O), molecular oxygen (O_2), Helium (He), and atomic Hydrogen (H). Global circulation causes interhemispheric transport of lighter species resulting in latitudinal variations in species and neutral density [11]. In addition to horizontal species movement, upwelling and downwelling can transport species vertically, impacting neutral density as a function of altitude. Horizontal and vertical transport mechanisms have been recently investigated by Sutton [12]. Certain species (e.g. nitric oxide (NO) and carbon dioxide (CO_2)) provide cooling mechanisms, particularly in response to geomagnetic activity [13, 14, 15].

At times, geomagnetic activity can dominate fluctuations in thermospheric mass density. During large storms, significant energy (and therefore heat) are induced into the thermosphere and can cause global increases in density by multiples of its pre-storm levels [16]. It is also one of the most difficult phenomena to model [17]. Geomagnetic storms originate from the Sun and propagate to Earth through complexly coupled systems leading to part of the difficulty in forecasting a solar event's impact on the thermosphere. A major issue in improving modeling capabilities during storms is the relative rarity of these events, leading to limited storm observations. Figure 2.1 shows how the disturbance storm time index (*Dst*) is distributed between December 1996 and April 2020. *Dst* will be explained in more detail later (Section 2.3.3), but generally, *Dst* values more negative than -75 *nT* denote a storm. This index is produced hourly, providing over 200,000 values across this time span, and only 1.34% of these values would characterize a geomagnetic storm.

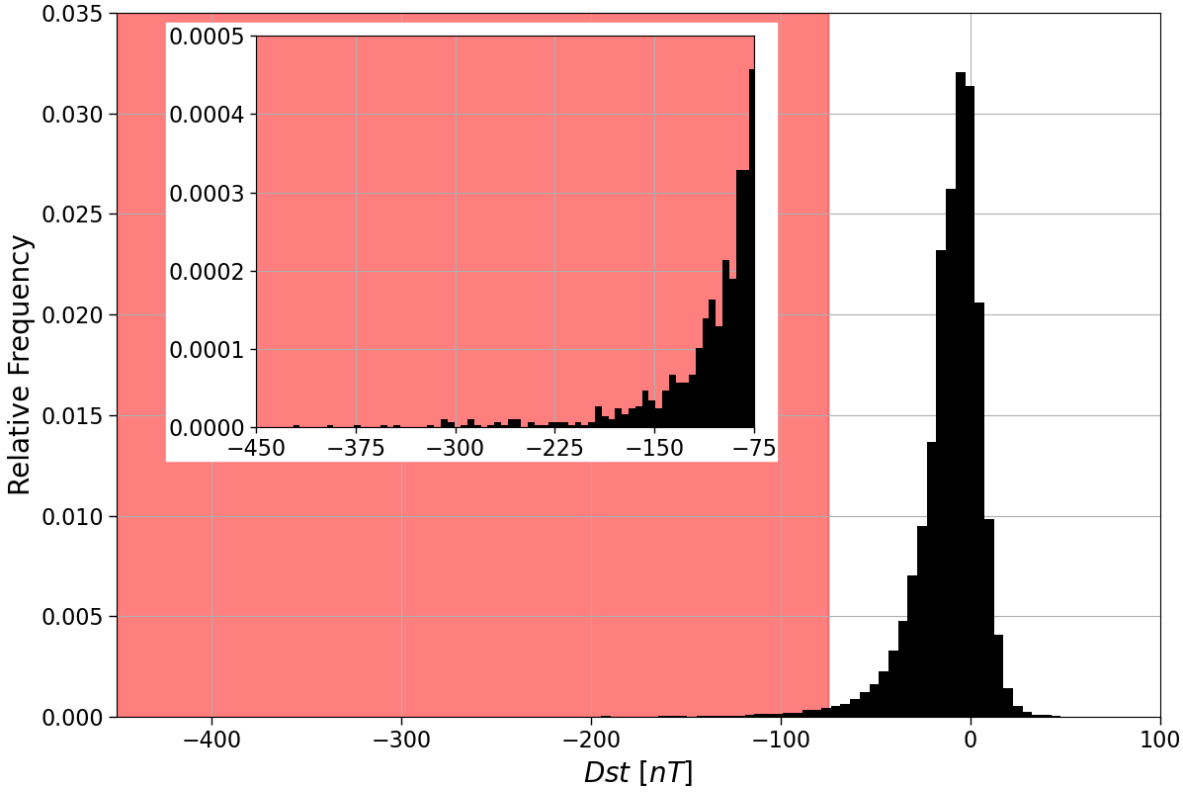

Figure 2.1: Distribution of *Dst* over two solar cycles with the shaded region denoting storm conditions. The secondary subplot shows the distribution focused on the storm conditions. Note: the subplot labels are consistent, but the axis limits are not.

2.2 Space Weather

Space weather is an amalgam of conditions and events occurring on the sun, in the resulting solar wind, and in the near-Earth geospace environment. Similar to terrestrial weather, many of these subsystems are highly coupled and difficult to predict. The consequences of space weather include but are not limited to: satellite charging, enhanced radiation, and geomagnetically induced currents (GICs) that can disrupt power grids. However, the space weather impact that will be the focus of this work is satellite drag in LEO. The proceeding subsections are meant to provide background knowledge on how activity from the Sun results in short-term disturbances to thermospheric mass density.

2.2.1 Solar Wind

In the Sun's inner core, nuclear reactions produce massive amounts of energy that radiate towards its surface. Once it reaches the tachocline, the boundary between the radiative and convective zones, convection takes over as the primary mode of energy transport towards the solar surface [18]. As the Sun is not a solid body, it has differential rotation which induces turbulence in the convection region and is linked to the dynamics and strength of its magnetic field [19]. The Sun's magnetic field has a strong dipole configuration, which reverses every eleven years, or one solar cycle [20] first observed by Babcock [21]. The magnetic field lines move radially outward from the Sun and remain connected even as the Sun rotates. This creates a swirling pattern known as the Parker spiral, seen in Figure 2.2 [22]. The solar wind is the continuous emission of plasma from the Sun along these magnetic field lines.

The solar wind is currently measured by the Advanced Composition Explorer (ACE) and Deep Space Climate Observatory (DSCOVR) satellites [23, 24] at the L_1 Lagrange point where the gravity of the Earth and Sun are equal. This allows both spacecraft to have a direct view of the Sun and mitigates the need for excess fuel to remain in the ideal location. Both spacecraft contain monitors that allow them to measure continuously varying Earth-bound solar wind characteristics (e.g. velocity, density, magnetic field strength). These are all crucial for models inside the geospace

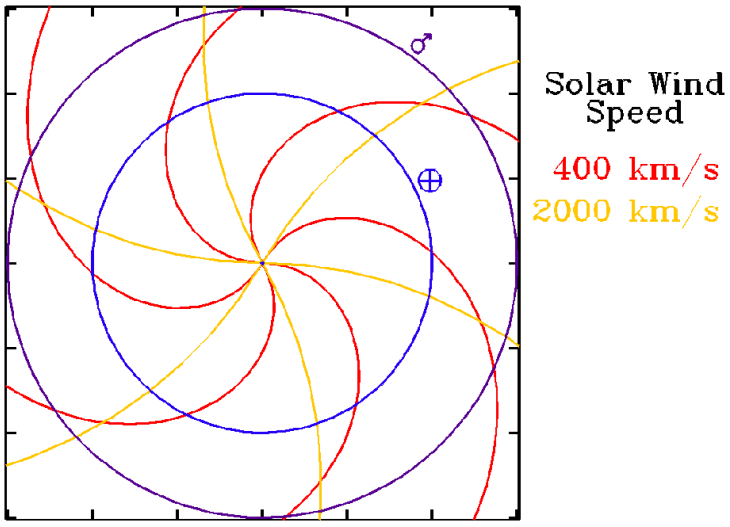

Figure 2.2: Parker spirals for different solar wind speeds. The orbits of Earth and Mars are shown in blue and purple, respectively. IC: NASA https://sdo.gsfc.nasa.gov/mission/spaceweather.php

system as well as for models that forecast the solar wind.

The solar wind has continuously-evolving properties that are dependent on the origin. Coronal mass ejections (CMEs) are massive bubbles of plasma that are ejected from the Sun, releasing a significant volume of solar wind [25]. The resulting solar wind also travels at velocities multiple times background levels. When this fast CME-driven solar wind interacts with slower solar wind, a shock is formed, and the sheath behind this contains compressed plasma and has increased magnetic field strength [26].

Coronal holes are another solar feature that can create stark changes in the solar wind. Coronal holes are portions of the Sun's corona that contain low-density plasma and have open magnetic field lines [27]. The solar wind emitted from coronal holes forms a high-speed stream (HSS) which has higher velocity than background solar wind but much less than that of a CME [28]. A unique property of coronal holes is that they are fairly persistent and evolve slowly. During solar minimum especially, HSSs tend to occur with a regular frequency of about 27 days, or a solar rotation.

2.2.2 Near-Earth Geospace Environment

The near-Earth geospace environment primarily consists of the magnetosphere, ionosphere, and thermosphere. As it pertains to space weather, the magnetosphere acts as a low-pass filter for rapid changes in the solar wind. Disturbances affect the Earth's magnetic field, and energy can then pass into the ionosphere-thermosphere system. These systems are all tightly-coupled.

2.2.2.1 Magnetosphere

The magnetosphere is a complex system shaped by Earth's magnetic field rooted in its outer core and externally shaped by the solar wind. Earth's magnetic field changes in space and time as does the solar wind that interacts with it. An artist's depiction of the magnetosphere is shown in Figure 2.3. In this figure, the left-most curved line represents the bow shock – due to the supersonic speed of the solar wind. The bright curve to the right of the bow shock is the magnetopause, or the boundary of the magnetosphere. The area between these two regions is referred to as the magnetosheath. The area enclosed by the magnetopause is the magnetosphere which is shown with yellow magnetic field lines.

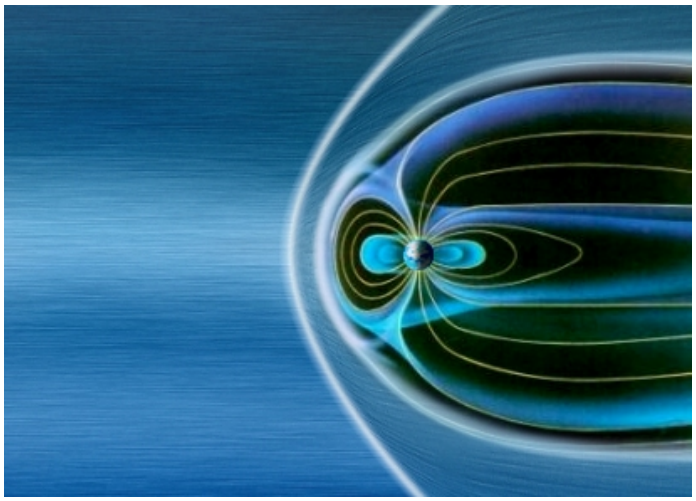

Figure 2.3: Depiction of the solar wind and the magnetosphere. IC: ESA/AOES Medialab https://www.esa.int/Enabling_Support/Operations/Rejigging_the_Cluster_quartet

Energy, mass, and momentum are transferred from the solar wind to the magnetosphere which creates geomagnetic activity [29]. Much of this coupling between these two systems is driven by magnetic reconnection. This is the process in which field lines of the magnetospheric plasma are connected to the solar wind wind on the dayside and move towards the magnetotail where they reconnect again; this process is known as the "Dungey cycle" [30]. The reconnection in the magnetotail can then allow energetic electrons to travel along field lines back towards Earth where they energize the upper atmosphere at the poles [31]. While there are numerous properties of the solar wind that play a role in the its interaction with the magnetosphere, the interplanetary magnetic field (IMF) having a strong southward component (B_Z) is critical for geoeffectiveness [32].

2.2.2.2 Ionosphere-Thermosphere System

The ionosphere and thermosphere are complexly coupled with the magnetosphere, and the combination is called the magnetosphere – ionosphere – thermosphere (MIT) system. A key distinction between the ionosphere and thermosphere is that they consist of charged and neutral particles, respectively. During geomagnetic storms, the energy induced into the auroral ionosphere, as a result of magnetic reconnection, causes ion-neutral collisions inducing energy into the system [33]. The energy deposition into the ionosphere and thermosphere through either Joule heating or particle precipitation causes both systems to heat and expand [34]. Although this energy enhancement occurs at high latitudes, they can travel to lower latitudes in the form of traveling ionospheric disturbances (TIDs) [35] and traveling atmospheric disturbances (TADs) [36].

2.3 Space Weather Modeling Efforts

Models have been developed for all the systems discussed thus far. Our understanding of space weather comes largely from observations as well as models that can adequately support observations. An example of this is a study done by Asplund et al. [37], where a solar model is refined to closely match observations, improving our understanding of the Sun's chemical composition. Physics-based space weather models can also be coupled to provide a realistic representation of the entire Sun-Earth system involving the feedback between subsystems [38]. Empirical models con-

sist of parametric equations that are fit to abundant observations made by various sensors. These are often used in an operational setting due to their computational simplicity relative to physical models.

2.3.1 Thermosphere Models and Data

Over the past six decades, the scientific community has developed and advanced thermospheric density models. A significant subset of these models are empirical. Even within this subset, there are multiple families/series of models that use different types of measurements and have evolved over decades. Three of these series, discussed by Emmert [8], are the MSIS [39], DTM [40], and Jacchia series [6]. Mass Spectrometer Incoherent Scatter Radar (MSIS) models typically use mass spectrometer and incoherent scatter radar measurements but have evolved and now incorporate additional data (e.g. accelerometer-derived density estimates). The Drag Temperature Model (DTM) series used orbit-derived density data but more recently incorporated accelerometer-derived density and mass spectrometer data. The Jacchia series of models (e.g. Jacchia-70 and the Jacchia-Bowman 2008 Empirical Thermospheric Density Model (JB2008)) strictly use both orbit- and accelerometer-derived density estimates. A major improvement in density modeling capabilities came with the introduction of real-time data assimilation. The High Accuracy Satellite Drag Model (HASDM) [41] is an assimilative model/framework that leverages JB2008 and Dynamic Calibration of the Atmosphere (DCA) to correct the density nowcast with satellite observations.

The availability of accelerometer-derived density estimates has been advantageous for model development and assessment. Over the lifetime of satellites with onboard accelerometers (e.g. CHAllenging Minisatellite Payload (CHAMP) and Gravity Recovery and Climate Experiment (GRACE)), we accumulate measurements over many altitudes and space weather conditions [42, 43]. Researchers have used these measurements to derive density estimates by removing accelerations from other sources (e.g. gravity and solar radiation pressure) [5, 44, 45, 46, 47, 48].

Physics-based thermosphere models play a significant role in scientific studies of the upper atmosphere. Three examples of these model types are Thermosphere-Ionosphere-Electrodynamics General Circulation Model (TIE-GCM) [49], Coupled Thermosphere Ionosphere Plasmasphere

Electrodynamic Model (CTIPe) [50], and Global Ionosphere-Thermosphere Model (GITM) [51]. While not true for all physics-based thermosphere models, TIE-GCM, CTIPe, and GITM all use the finite difference method to solve the physical equations, couple the ionosphere, and generate self-consistent electric fields at low to middle latitudes [8]. These models also have varying upper boundaries dependent on solar conditions and the corresponding pressure levels. The thermosphere models and datasets used in this work are outlined in the following sections.

2.3.1.1 Mass Spectrometer and Incoherent Scatter Radar Series

The Naval Research Laboratory Mass Spectrometer and Incoherent Scatter radar (NRLMSIS 2.0) empirical thermosphere model [52] is the most recent version of MSIS models dating back to the original MSIS-86 model [53]. The data sources for its predecessor (NRLMSISE-00) are observed satellite drag, accelerometer data, incoherent scatter radar T_∞ and lower thermosphere temperature data, and occultation-derived O_2 density data [39]. NRLMSIS 2.0 sought to improve the definition of composition/structure at low altitudes (below 100 km) by incorporating recent data from low-altitude temperature measurements. It also assimilated additional atomic oxygen and hydrogen measurements to overcome previous limitations.

NRLMSIS 2.0 uses the *ap* index to account for geomagnetic activity. There are two *ap* options when running NRLMSIS 2.0: use only the daily average (known as *Ap*) and current 3-hour value, or use a time history of the index. This time history includes *Ap*, current *ap*, ap_3, ap_6, ap_9, ap_{12-33}, and ap_{36-57}. The single numerical subscripts refer to the value of the index that many hours prior to the epoch. The combination of two numbers in the subscript refers to the average value over that many hours prior to the epoch (e.g. ap_{12-33} is the average *ap* value from 12 to 33 hours prior to the epoch). This nomenclature for geomagnetic drivers will be used throughout this dissertation.

2.3.1.2 High Accuracy Satellite Drag Model

The most recent in the Jacchia series is the JB2008 density model. JB2008 was an improvement to its predecessors and incorporated new solar and geomagnetic indices to drive the model. It uses the F_{10}, S_{10}, M_{10}, and Y_{10} indices and proxies to model variations caused by solar heat-

ing. In addition to *ap*, JB2008 utilizes *Dst* to improve model density during geomagnetic storms. These indices are used in temperature corrections, semiannual functions, and new *Dst* temperature equations. The model reduced non-storm density errors by more than 5% and reduced storm-time density errors from Jacchia-70 by more than 60%, from NRLMSISE-00 by more than 35% and from JB2008 (with only *ap*) by 16% [6]. These drivers are described in Section 2.3.3.

HASDM is an assimilative framework using JB2008 as a background density model. HASDM improves upon the density correction work of Marcos et al. [54] and Nazerenko et al. [55] 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 [56]. The HASDM algorithm uses a prediction filter that employs wavelet and Fourier analysis for the correction coefficients [41]. Another highlight of HASDM's novel framework is its segmented solution for the ballistic coefficient (SSB). 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 2000-2020 make up the SET HASDM density database, upon which this work is based. The database contains density predictions with 15° longitude, 10° latitude, and 25 km altitude increments spanning from 175 - 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. [41], and for details on SET's validation process and on the database, the reader is referred to Tobiska et al. [57].

2.3.1.3 Thermosphere-Ionosphere- Electrodynamics General Circulation Model

The Thermoshere-Ionosphere- Electrodynamics General Circulation Model (TIE-GCM) is part of the Thermosphere General Circulation Model (TGCM) series dating back to 1981 [58]. TIE-GCM was developed by Richmond et al. [59] and managed to incorporate electrodynamic interactions between the thermosphere and ionosphere systems. As stated by Qian et al. [49], "TIE-GCM self-consistently solves the fully coupled, nonlinear, hydrodynamic, thermodynamic, and conti-

nuity equations of the neutral gas, the ion and electron energy and momentum equations, the ion continuity equation, and neutral wind dynamo". TIE-GCM is important for studying different phenomena in the thermosphere due to its ability to provide a constrained dynamic evolution of the system. However, the high computational cost and need for significant parallelization limits its application to operations or collision assessment.

2.3.1.4 Satellite Accelerometer Data

The CHAMP and GRACE density datasets used in this work (for either model development or validation) are from Mehta et al. [48], which originate from Sutton [46] but are scaled to account for higher fidelity satellite geometry and improved gas-surface interaction simulations [60, 61, 62]. However, there is no correction to the solar radiation pressure accelerations which used the simplified 13-panel geometry. Both satellites have near-polar orbits, covering nearly all latitudes, and over their respective lifetimes, CHAMP and GRACE datasets cover altitudes ranging from 300–535 km. This, in conjunction with the date range covered by the satellites, makes their density estimates invaluable for model comparison. Figure 2.4 shows the altitudes each dataset covers along with orbit-averaged densities over their mission spans.

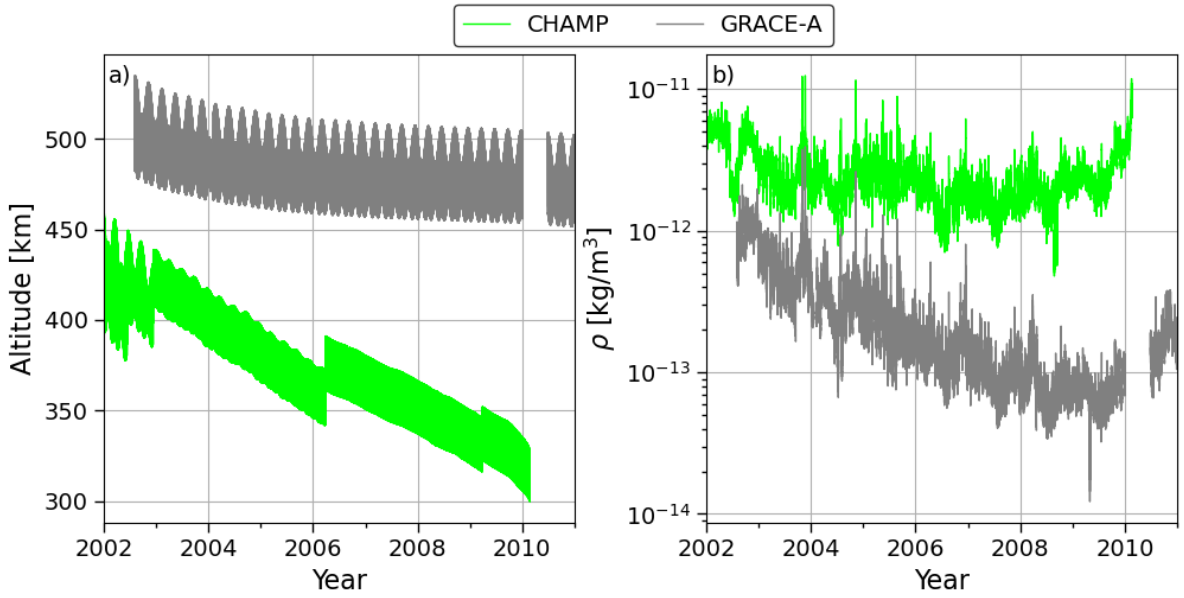

Figure 2.4: Altitude (a) and orbit-averaged densities (b) for CHAMP and GRACE-A.

The orbit-averaged densities were computed using a centered window with a span of 90 minutes, approximately one orbit. Discontinuities in Figure 2.4 represent data gaps. Both the CHAMP and GRACE-A datasets contain files for every day containing information such as GPS time, local solar time (LST), latitude, altitude and density. CHAMP has measurements every 10 seconds, while GRACE-A provides measurements every 5 seconds.

2.3.2 Exospheric Temperature

As with many models (e.g. DTM and JB2008), MSIS heavily relies on temperature profiles to determine species densities and therefore mass density throughout the thermosphere. A key parameter in predicting the temperature profile is the exospheric temperature (T_∞) which is the asymptotic value that the temperature profile approaches at the top of the thermosphere, or thermopause [63, 64]. MSIS uses the Bates-Walker temperature profile [65]. Figure 2.5 shows MSIS temperatures from the ground to 800 km. The temperature dependence on solar activity is prominent, as is the difference between day and night. For all four temperature profiles, the variation slows above 250-300 km and becomes relatively constant. This temperature "limit" for each curve corresponds to the MSIS T_∞ prediction for that condition.

2.3.2.1 Exospheric Temperature Estimates

Typically, NRLMSIS 2.0 predicts the exospheric temperature as a function of position and space weather drivers. Using this computed value, the model then calculates species densities as a function of altitude and therefore neutral mass density. The user can override the internally computed T_∞ and MSIS will determine density based on the provided value. In the past, this has been leveraged in numerical schemes to match MSIS to satellite measurements in order to estimate T_∞ [66, 67, 68]. Weimer et al. [69, 70] used the CHAMP and GRACE density estimates described in Section 2.3.1.4 from 2001 through 2010 to perform a similar derivation of exospheric temperatures. They also used additional satellite data from the Swarm missions [71]. They had varied T_∞ in NRLMSIS 2.0 using a binary search method to match satellite density until the temperature was determined to 2 K.

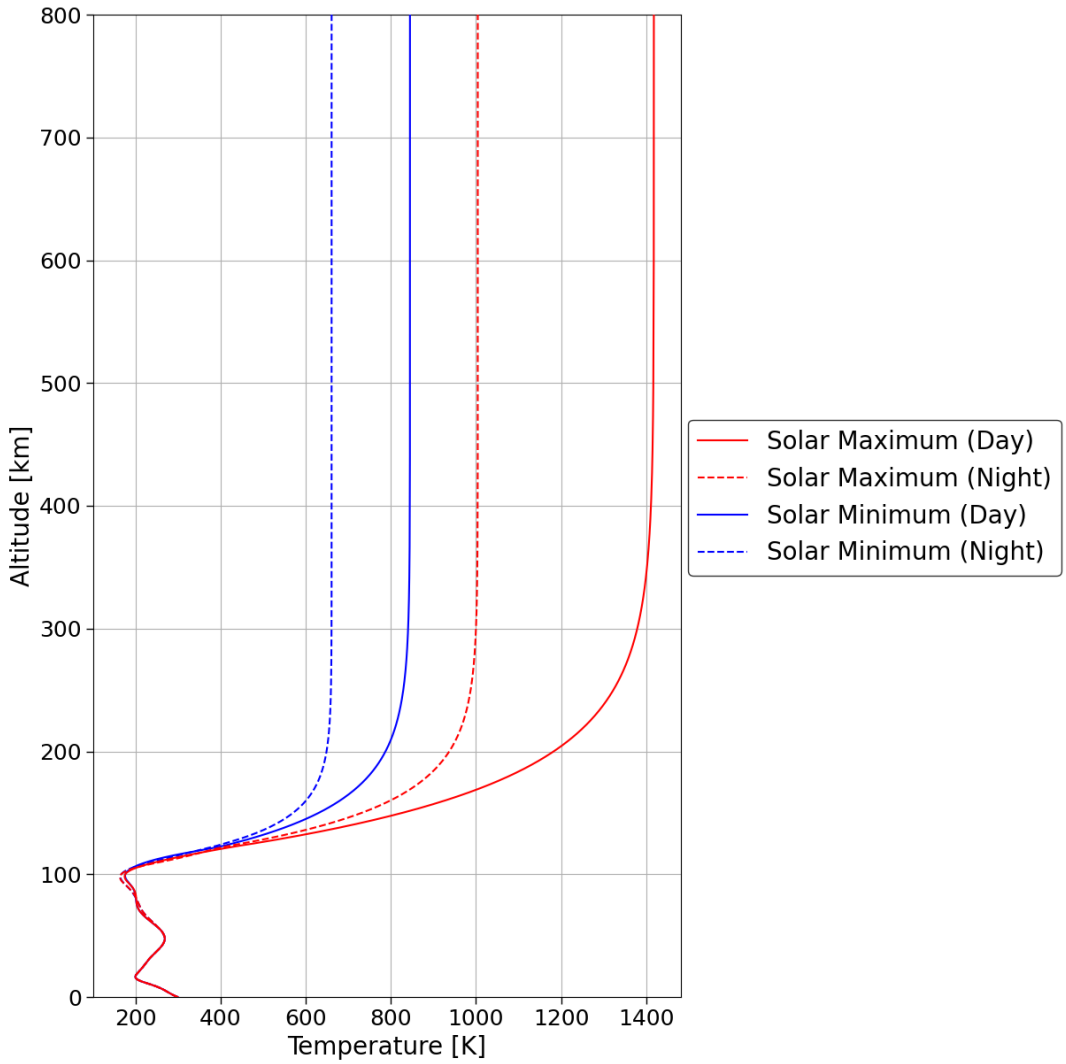

Figure 2.5: Temperature profiles using MSIS for solar maximum and solar minimum conditions.

The density estimates originate from the following sources – CHAMP 2001: Doornbos [5], CHAMP 2002-2010: Mehta et al. [48], GRACE 2002-2009: Mehta et al. [48], GRACE 2010: Sutton [46], Swarm A 2013-2018: Astafyeva et al. [72], and Swarm B 2012-2016: Astafyeva et al. [72]. Note that only GRACE-A measurements are used due to the similarity of the GRACE-A and GRACE-B orbits. The CHAMP and GRACE density estimates originate from accelerometer measurements and span an altitude range of 300–535 km while the Swarm A and B density estimates are obtained from GPS data and span an altitude range of 437–545 km. The cadence of the satellite estimates are 10 s, 5 s, 30 s, and 30 s for CHAMP, GRACE, Swarm A, and Swarm B,

respectively. There are over 82 million samples total for model development and evaluation.

EXospheric TEMperatures on a PoLyhedrAl gRid (EXTEMPLAR) – an exospheric temperature model that can be integrated with MSIS – is developed based on a polyhedral grid made of 1,620 cells [69, 70]. The measurements from CHAMP, GRACE, Swarm A, and Swarm B were binned to the closest grid cell for model development. Figure 2.6 shows the distribution of measurements across the polyhedral grid. This gives a sense of the spatial distribution of the satellite density and exospheric temperature estimates. They are most heavily distributed at the poles due to the satellites' high inclination.

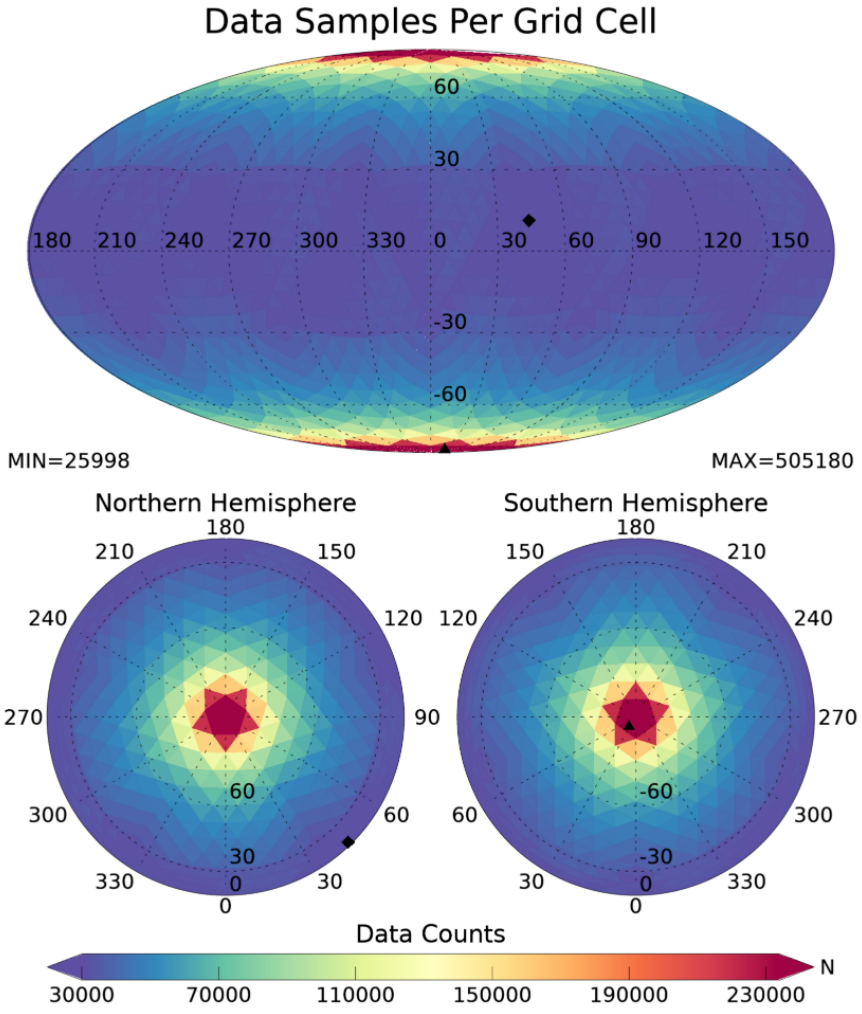

Figure 2.6: Number of samples for each of the 1,620 polyhedral grid cells.

2.3.3 Model Drivers

The JB2008 model, for example, uses global exospheric temperature equations driven by four solar indices/proxies to represent different solar heating sources [73, 6]. From ISO 21348 [74], an index is a measured indicator of level of activity while a proxy is a surrogate for other physical processes. The four solar indices and proxies are all report in solar flux units (10^{-22} W m^{-2} Hz^{-1}) which are denoted as sfu.

The $F_{10.7}$ proxy (referred to here as F_{10}) has a strong correlation to solar EUV irradiance which has led to its long-time use as a surrogate for solar EUV energy. However, F_{10} has no physical relation to solar EUV irradiances. S_{10} is an index indicative of activity of the integrated 26-34 nm bandpass solar chromospheric EUV emission, which penetrates to the middle thermosphere and is absorbed by atomic oxygen. The M_{10} proxy is used as a surrogate for FUV photospheric 160 nm Schumann-Runge Continuum emissions, which penetrate to the lower thermosphere and cause molecular oxygen dissociation. The fourth solar index is Y_{10}. This is a hybrid index of solar coronal 0.1-0.8 nm X-ray emissions and 121.6 nm Lyman-alpha, both of which penetrate to the mesosphere and participate in water chemistry. The S_{10}, M_{10}, and Y_{10} indices and proxies were derived from actual solar irradiance measurements and scaled to F_{10} magnitudes in the original JB2008. This has also allowed an ease of comparison between these disparate time series.

To capture the impact of geomagnetic activity, JB2008 uses a synthesis of *ap* and *Dst* indices. The *ap* index is a measure of global geomagnetic activity derived from twelve observatories that fall between 48° N and 63° S in latitude [75]. The utilization of *ap* during quiet geomagnetic conditions results in low density errors, but *Dst* proves to be a more effective driver during storm times [6]. *Dst* is an index that represents the strength of the storm-time ring current in the inner-magnetosphere [73]. For further details on all of the JB2008 drivers, the reader is referred to Tobiska et al. [73] and ISO 14222 [76].

While JB2008 uses *ap* and *Dst* for geomagnetic storm characterization, other indices are used by the modeling community. *Kp* is a legacy driver for global geomagnetic activity. It is strongly related to *ap* but is quasi-logarithmic. Both planetary indices have 28 discrete values: *ap* ranges

from 0 to 400 and *Kp* ranges from 0 to 9. *Kp* values are presented as integers with plus and minus indicators that denote ± 1/3 from the integer value. Table 2.1 and Figure 2.7 show the corresponding values of the two indices. If interpolation between indices is required outside of the discrete value shown in the table (e.g. for modeling purposes), cubic spline interpolation has been shown to be the appropriate method [75].

Table 2.1: *Kp* and *ap* discrete values. The second *Kp* row shows the values in numerical format up to two decimal places [75].

Kp	0o	0+	1-	1o	1+	2-	2o	2+	3-	3o	3+	4-	4o	4+
Kp	0.00	0.33	0.67	1.00	1.33	1.67	2.00	2.33	2.67	3.00	3.33	3.67	4.00	4.33
ap	0	2	3	4	5	6	7	9	12	15	18	22	27	32
Kp	5-	5o	5+	6-	6o	6+	7-	7o	7+	8-	8o	8+	9-	9o
Kp	4.67	5.00	5.33	5.67	6.00	6.33	6.67	7.00	7.33	7.67	8.00	8.33	8.67	9.00
ap	39	48	56	67	80	94	111	132	154	179	207	236	300	400

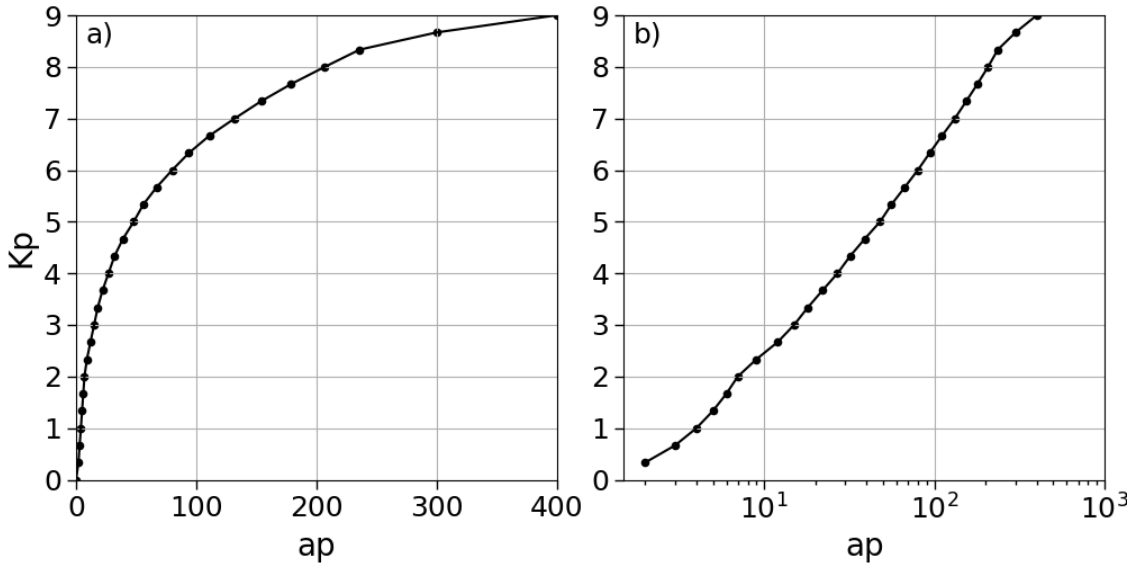

Figure 2.7: *Kp* vs *ap* relationship in linear (a) and semi-log (b) presentation.

2.3.3.1 Additional Model Drivers

EXTEMPLAR uses different geomagnetic drivers than most empirical models, particularly S_N, S_S, ΔT. The two S inputs are Poynting flux totals in the Northern and Southern hemispheres [77, 78]. ΔT is a parameter derived by Weimer et al. [69, 70] and represents exospheric temperature perturbations; it is a function of Poynting flux and simulated nitric oxide emissions. While *Dst* has been previously discussed as an improvement over *ap*/*Kp*, it has its own shortcomings. It has a one-hour cadence, which is three times higher than that of *ap*/*Kp*, but it is still too coarse for models like EXTEMPLAR which are based on measurements with cadences as high as five-seconds. *SYM-H*, the longitudinally symmetric component of the magnetic field disturbances, has similar characteristics to *Dst*. A key benefit over *Dst* is its 1-min cadence [79].

Common temporal inputs for density models are universal time (UT) and day of year (doy). An issue when building a regression model with these drivers is the discontinuity at the end of the day/year. To overcome this, they can be transformed to become continuous about the boundaries. The transformed time inputs $t_1 - t_4$ are defined in Equation 2.1.

$$t_1 = sin\left(\frac{2\pi doy}{365.25}\right) \quad t_2 = cos\left(\frac{2\pi doy}{365.25}\right) \quad t_3 = sin\left(\frac{2\pi UT}{24}\right) \quad t_4 = cos\left(\frac{2\pi UT}{24}\right) \tag{2.1}$$

Typical spatial coordinates for the upper atmosphere are longitude and latitude. Longitude is not ideal for studying and modeling the atmosphere as the orientation of its dominant diurnal structure is changing with UT. Local solar time (LST) is a better longitudinal coordinate to use since it is based on the location of the sun. This is used throughout all modeling efforts as either an input or to define the gridded data. Like longitude and the temporal inputs, its linear structure causes discontinuities about midnight. This is solved with a similar transformation shown in Equation 2.2.

$$LST_1 = sin\left(\frac{2\pi LST}{24}\right) \quad LST_2 = cos\left(\frac{2\pi LST}{24}\right) \tag{2.2}$$

Chapter 3. Machine Learning Background

A neural network is a collection of computational cells (or neurons) connected in some form through multiplicative connections (or weights). Neural networks were first conceived by McCulloch and Pitts [80] when they described a computational representation of brain neurons and synapses with calculus and statistical theory. In the late 1950s, the first artificial neural network (ANN) was developed and is known as the perceptron [81]. Backpropagation is the process in which the network parameters are updated based on observations and was fundamental to the development of modern neural networks [82, 83]. Another significant advancement in neural networks was the implementation of graphics processing units (GPUs) for vastly increased training speeds in convolutional neural networks (CNNs) [84]. The application of neural networks and ML has expanded in the decades since its inception and has found its way into all industries from medicine to game playing.

Machine learning (ML) is a subset of artificial intelligence where the internal parameters of a neural network are learned in an iterative process, similar to the way humans learn. It has been growing in popularity in the space weather community in recent years. The following chapter is focused on the development of ML thermosphere models, so we focus on the basics here to provide the necessary background knowledge for the remainder of this work.

3.1 Neural Network Terminology

To cover the basics of neural networks, we consider the simple model displayed in Figure 3.1. The main characteristics of a model will be decided by the scope of the problem. This model is developed for a problem with two input dimensions (x,y) and one output dimension (z). This could be for predicting temperature (z) as a function of longitude and latitude (x and y).

This is representative of a fully-connected model where weights connect all nodes (also known as neurons or units) between layers. This model also has bias units. This is why although we only have two inputs, the input layer has three units. When you provide x and y, the bottom two nodes

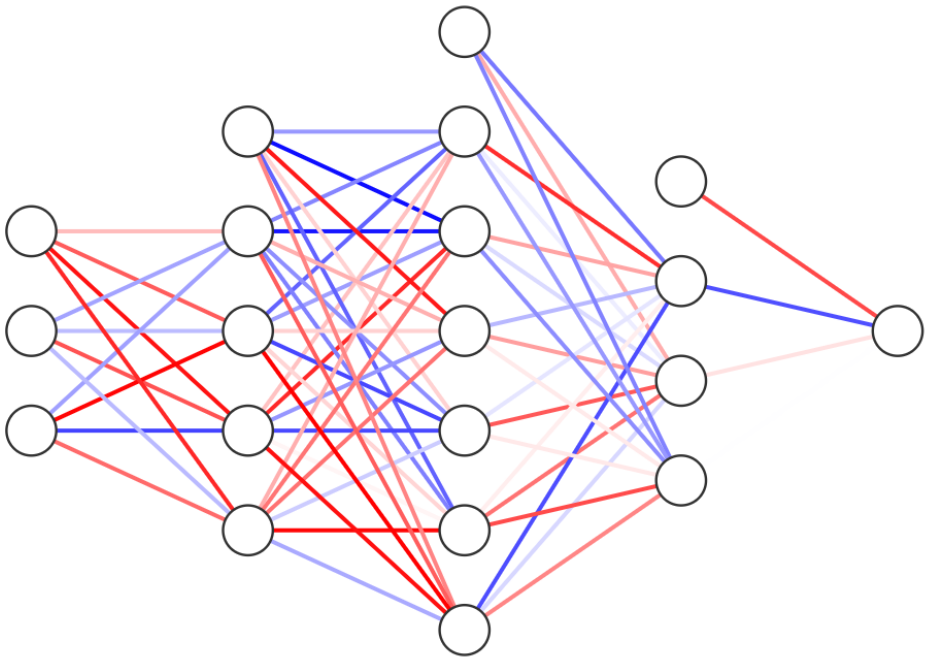

Figure 3.1: Diagram of an ANN with bias units. The colors and opaqueness denote the sign and magnitude of the random weights, respectively (via: https://alexlenail.me/NN-SVG/). Note: the flow of information is left-to-right.

in the input layer would have these values connected to the associated weights. Bias units always have the value 1.0 and the weights are therefore determining the bias to each downstream node. Since the value in the bias nodes are fixed, there are no upstream weights connecting to them.

There appear to be five total layers in this network. The first is the input layer with two defined inputs. The final layer is the output layer with one node (z). In between, there are three *hidden layers*. The flow of information through the model works as follows. The inputs $\mathbf{x}$ are multiplied with the weight matrix (W_1) between the input layer and the first hidden layer. This will result in four values, one for each node in the first hidden layer (not considering the bias). These values are different weighted sums of the upstream nodes. They all go through an activation function (explained later) chosen for this layer. The process is now repeated to move from the first to the second hidden layers and so on until an output value is computed.

3.1.1 Activation Functions

Activation functions are user-defined functions attributed to each layer or node in a neural network that can determine (a) whether or not a node is "activated" and/or (b) how much nonlinearity is being added to the system. The weighted sum going into a node (to be referred to as ϕ) will be passed as an input to an activation function f. The output of the node will then be $f(\phi)$. Some common activation functions are displayed in Figure 3.2. The functional forms can be found in Table 3.1.

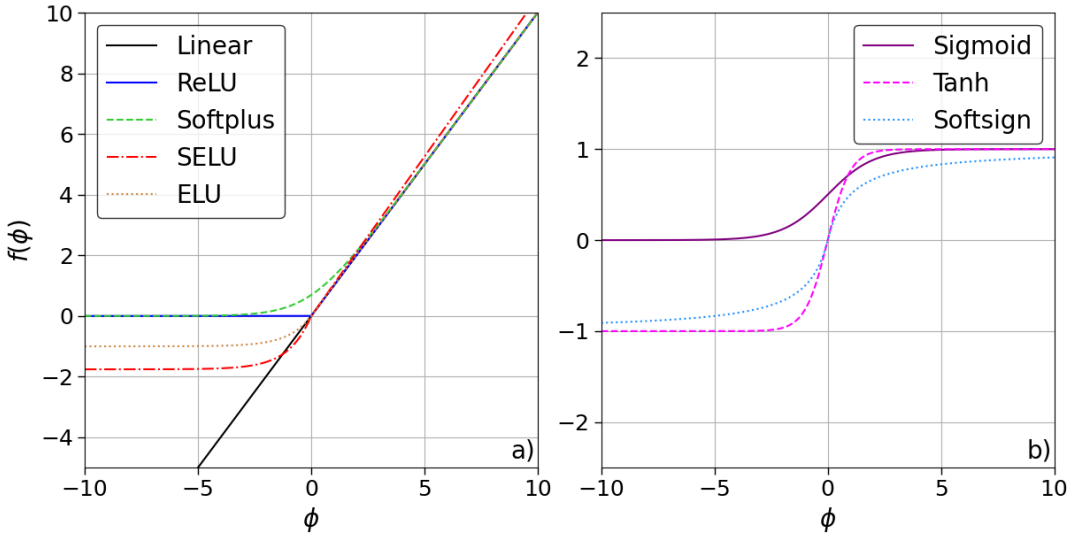

Figure 3.2: $f(\phi)$ for un/partially-bounded (a) and fully-bounded (b) functions.

The linear activation function will simply pass the weighted sum of the previous layers inputs. Linear activation functions are not typically useful as hidden layer activations, because their output is a linear combination of the upstream layer. If all hidden layers use the linear activation, the output will be a linear combination of the inputs, regardless of the number of layers. However, they are sometimes used to make a linear projection (e.g. in a generative adversarial network). A linear activation can also be useful as an output activation if the model output is fully unbounded. As seen in Figure 3.2, the linear function is the only one that is fully unbounded that is used in

Table 3.1: Functional form of all activation function in Figure 3.2.

Name	**Function, $f(\phi)$**
Linear	$f(\phi) = \phi$
ReLU	$f(\phi) = max(0, \phi)$
ELU	$f(\phi) = \phi$ if $\phi > 0$ $f(\phi) = \alpha\left(e^{\phi} - 1\right)$ if $\phi \leq 0$
SELU	$f(\phi) = \lambda\phi$ if $\phi \geq 0$ $f(\phi) = \alpha\lambda\left(e^{\phi} - 1\right)$ if $\phi < 0$
Softplus	$f(\phi) = ln\left(1 + e^{\phi}\right)$
Tanh	$f(\phi) = \frac{e^{\phi} - e^{-\phi}}{e^{\phi} + e^{-\phi}}$
Sigmoid	$f(\phi) = \frac{1}{e^{\phi} + e^{-\phi}}$
Softsign	$f(\phi) = \frac{\phi}{1 + \lvert\phi\rvert}$

this work. The only other commonly used unbounded activation is parametric rectified linear unit (PReLU) which takes the form $\alpha\phi$ when $\phi < 0$ and ϕ when $\phi \geq 0$ [85]. Another form of PReLU is called leaky rectified linear unit where $\alpha = 0.01$. These functions are not used in this work, because linear functions are used only as output activations to perform an unbounded transformation.

The positively unbounded activations (Figure 3.2 (a)) used in this work are rectified linear unit (ReLU), exponential linear unit (ELU), scaled ELU (SELU), and softplus (all defined in Table 3.1). The purpose of using these in the hidden layers is to add nonlinearity to the model. Choosing the most appropriate one requires a trial-and-error approach. The ReLU activation can also choose whether or not a neuron is activated. If the output of a neuron is less than zero, the activation outputs zero, essentially nullifying the node entirely. Both ELU and SELU have parameters that can be modified. In Table 3.1, the α for ELU has a default value of 1.0, and the default values for α and λ in the SELU function are 1.6733 and 1.0507, respectively.

The fully-bounded activation functions (Figure 3.2 (b)) are hyperbolic tangent (tanh), sigmoid, and softsign. Tanh and softsign are both bounded between -1.0 and 1.0, but softsign has a smoother

gradient. However, tanh is more computationally efficient due to its derivative: $f'(\phi) = 1 - f(\phi)^2$. Sigmoid, also known as the logistic function, is bounded between 0.0 and 1.0. Its derivative is also computationally efficient to compute, $f'(\phi) = f(\phi)(1 - f(\phi))$.

3.1.2 Loss Functions

Loss functions – or objective functions – inform a neural network of its goal. If the goal is to predict data with low errors, then a mean absolute error (MAE) or mean square error (MSE) activation function would work best. The MSE function is defined as

$$MSE(z, \hat{z}) = \frac{1}{n} \sum_{i=1}^{n} (z - \hat{z})^2 \tag{3.1}$$

where z is the true value, $\hat{z}$ is the model prediction, and n is the batch size. The goal with MSE is to minimize the loss value across the training dataset. This brings up the concept of training in batches (as signified by n in Equation 3.1). In training, the model will make predictions with the supplied inputs and compute a loss based on the predicted value. This loss is used to compute gradients and update the model's weights using backpropagation. A batch size is the number of samples to be passed through the model before computing an average loss and updating weights. This can help generalize the model, and the lower frequency of weight updates speeds up the training process.

Other common loss functions are used in machine learning, depending on the application. However, they will not be discussed as they are not directly relevant to this work. Two other loss functions will be used for uncertainty quantification and will be discussed in Sections 4.1.1.2 and 4.2. Given a chosen loss function (consider MSE), there will be a loss surface or manifold that contains the loss associated with every weight in the neural network. The manifold is n_w-dimensional where n_w is the number of weights in the model. This can only be visualized for one or two weights, as shown in Figure 3.3.

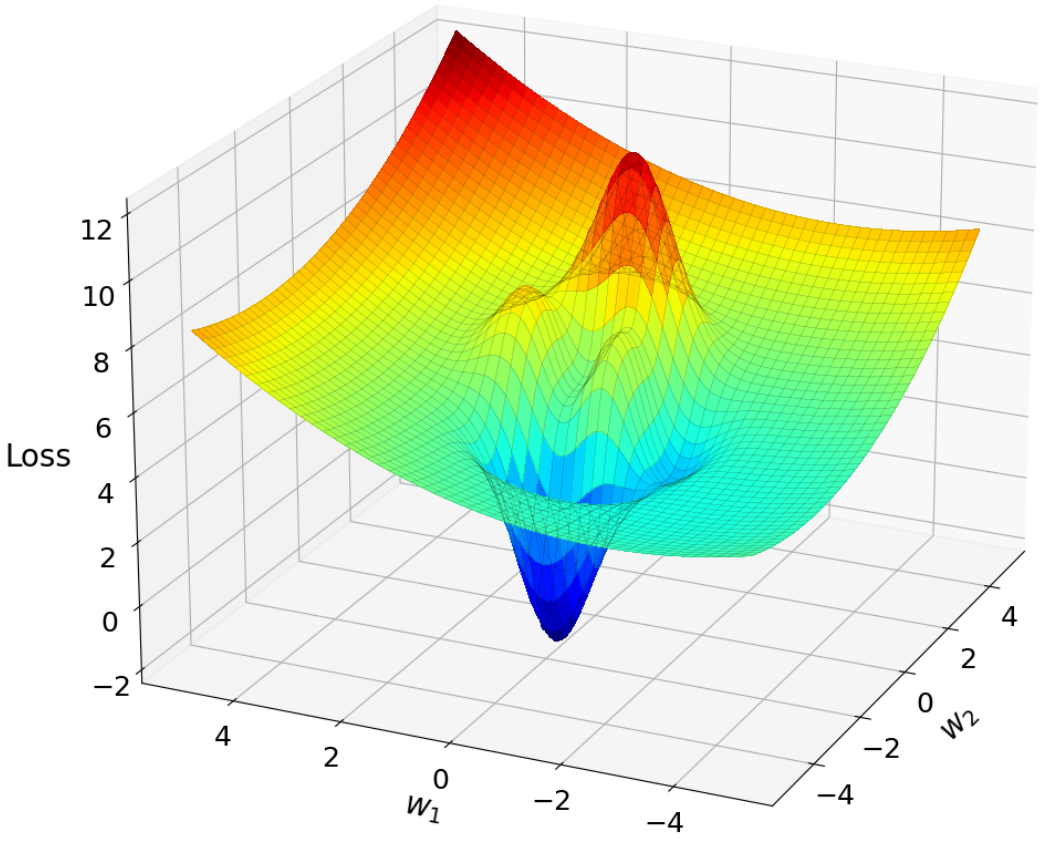

Figure 3.3: Example of a loss manifold for a simple neural network with two weights.

3.1.3 Optimizers

The goal in training is to find the minimum loss value, denoted by the darkest shade of blue in Figure 3.3. Since weights are randomly initialized, the model would start at a random point on this surface. How the model traverses this surface is determined by the optimizer. The simplest optimizer is standard gradient descent. This simply computes the gradient at a particular location and moves in the opposite direction by an amount determined by both the gradient and the learning rate (commonly denoted by η). Updating weights using this method is achieved by computing

$$w_{t+1} = w_t - \eta \frac{\partial \mathcal{L}}{\partial w_t} \tag{3.2}$$

where w_{t+1} are the updated weights, w_t are the current weights, and $\mathcal{L}$ is the loss. This simple way to train a neural network is effective, but other optimizer algorithms have been developed to make the training process both faster and more robust. While they will not be thoroughly explained, the following optimizers are used in this work because of their improvements with momentum and/or adaptive learning rate: Adaptive Moment Estimation (Adam) [86], Adam with Nesterov momentum (Nadam) [87], Adaptive Gradient Algorithm (Adagrad) [88], Adadelta [89], and Root Mean Squared Propagation (RMSProp) [90].

3.1.4 Normalization

Typically, inputs and outputs for a neural network will have wide ranges of values and can have large magnitudes. TensorFlow randomly initializes the weights in each layer from a uniform distribution that cannot have an initial magnitude larger than $\sqrt{3}$ (https://keras.io/api/layers/initializers/#glorotuniform-class). Having a large discrepancy in the magnitude and range of the input/output variables and the weights can lead to longer training times and poor performance [91]. This creates a need for data preprocessing and specifically normalization. This can be achieved with standard normalization,

$$\widetilde{\theta} = \frac{\theta - \text{mean}(\theta)}{\text{standard deviation}(\theta)} \tag{3.3}$$

where θ is an input/output variable, and $\widetilde{\theta}$ is the normalized θ. This ensures that all all input and output variables to the model will have the same first two moments as a standard normal distribution. This operation can be easily reversed to obtain un-normalized model predictions.

3.2 Hyperparameter Tuning

Hyperparameters are all of the topics discussed in Section 3.1. If hyperparameters are randomly selected, it is unlikely that the model will perform to proper standards after training. Furthermore, experience with ML may better inform you to choose a set of hyperparameters resulting in a good model, but there are still too many variations to land at a near-optimal solution. A standard ap-

proach to hyperparameter selection is systematically testing choices for each hyperparameter. This is time-consuming and not nearly exhaustive enough. Hyperparameter tuning is an advancement in ML that eases many of these aforementioned problems.

Keras Tuner [92] is a hyperparameter tuner developed for Tensorflow and Keras ML libraries in Python. The user can choose a tuner scheme (e.g. random search or Bayesian optimization) and ranges/choices for each hyperparameter. These choices can include the number of hidden layers, the number of neurons in each layer, activation functions for each layer, dropout rates, and optimizer. The user can will also choose the objective, number of trials, and executions per trial. The objective would be the metric that will be either minimized or maximized. In this work, minimizing the validation loss is the objective. The number of trials refers to the number of model architectures to be tested before the tuner is completed. Executions per trial deals with weight initialization (see Section 3.1.3). The number of executions per trial will be the number of times a model is trained with re-initialized weights for a given architecture/trial.

When using the Bayesian optimization tuner scheme, there is an option for the number of initial points (n_{ip}). This is due to the two-stage approach used by the scheme. In the first n_{ip} trials, the architectures will be randomly selected from the user-defined search space. Once these random trials are complete, the tuner uses a Gaussian process model to choose all future architectures. This GP model will ingest all data from previous trials to associate each hyperparameter to model performance. The end goal is to identify an architecture with near-optimal performance given the dataset and goal (or objective).

3.3 Toy Problem

Everything discussed in this chapter is combined to develop a neural network using a toy dataset for global temperature. This dataset was created by converting global density data using a constant scalar value to vary between 20 and 90°F. The dataset has fixed intervals of 15° longitude and 20 latitude values evenly spaced between -90 and 90°. There are also 24 separate snapshots at different universal times. Therefore, the dataset is of the shape $24 \times 24 \times 20$ in time $\times$ longitude $\times$ latitude.

This was reshaped so there were $24 \cdot 24 \cdot 20 = 11,520$ samples. Longitude, latitude, and time are the only values used as model inputs since this is a simple exercise. The input and output data is normalized per Equation 3.3. A tuner is then defined with data ranges and choices shown in Table 3.2

Table 3.2: Hyperparameter tuner parameters (left) and search space (right) for the toy temperature problem.

Tuner Option	**Choice**	**Parameter**	**Values/Range**
Scheme	Bayesian Optimization	*Number of Hidden Layers*	1–10
Total Trials	50	*Neurons*	min = 16, max = 512, step = 4
Initial Points	25	*Activations*	relu, softplus, tanh, sigmoid, softsign, selu, elu
Repeats per Trial	2	*Dropout*	min = 0.01, max = 0.60, step = 0.01
Minimization Parameter	Validation Loss	*Optimizer*	RMSprop, Adam, Adadelta, Adagrad

The objective of the tuner is to minimize the MSE on the validation set. In a real application, splitting the data into training, validation, and test sets is an important consideration and must be done properly. For each application considered in this work, a thorough explanation will be provided. However, this demonstration uses 10,000 random samples for training and the remaining 1,520 for validation in an effort to keep this simple. The tuner is then run with each model being trained for 50 epochs with a batch size of 128. The best model has the following architecture. The first hidden layer has 312 neurons, the ReLU activation, and a dropout rate (see Section 3.6) of 3%. The second hidden layer has 200 neurons, the ReLU activation, and a dropout rate of 44%. The model was trained with the Adam optimizer. This model was then used to predict a global temperature map for two times: 12:00 UT and 04:00 UT. Its predictions are compared to the ground truth in Figure 3.4.

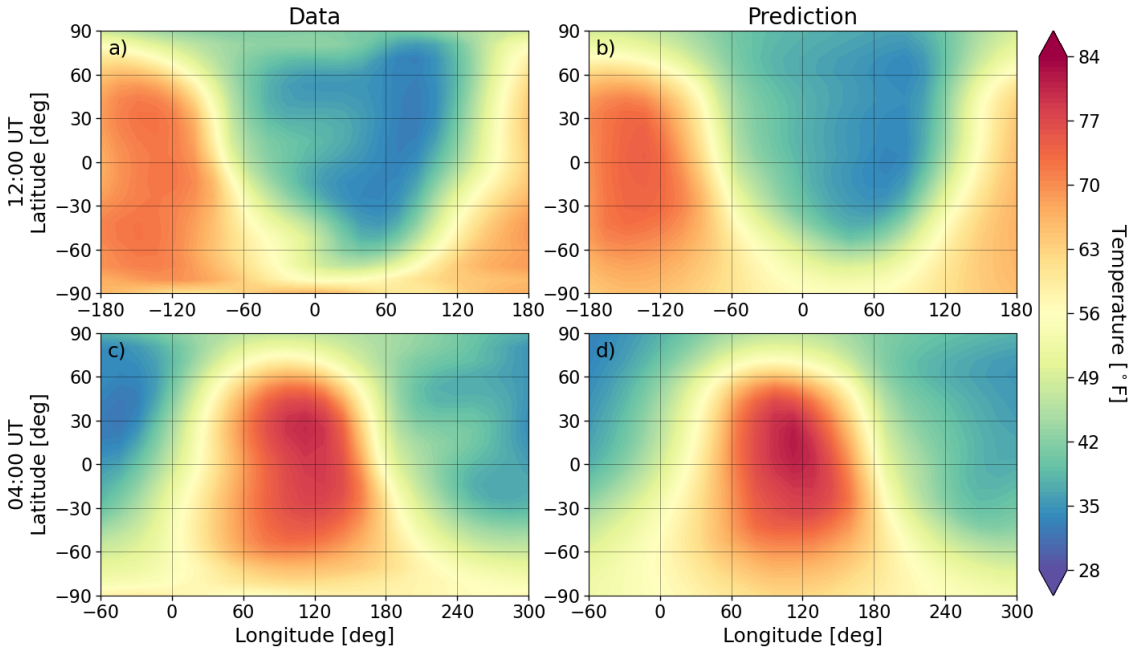

Figure 3.4: Original temperature data (a,c), and prediction (b,d) for 12:00 UT (a,b) and 04:00 UT (c,d).

Given the simple dataset, the model was able to adequately learn the relationship between location, time, and temperature. The general structures are indiscernible between the data and prediction with the exception of small-scale features on the night-side. Keep in mind the simplicity of this problem and dataset. In the remaining applications in this work, careful consideration is required when it comes to feature selection, data splitting, data preparation, and model evaluation.

3.4 Long Short-Term Memory Neural Networks

The previously described neural networks simply learn the relationship between inputs and outputs for a given time step; they are static models. For recurrent neural networks (RNNs), the corresponding inputs and outputs are concatenated, and the inputs for a given time-step are these stacked input and output combination. The number of previous time steps used is a new hyperparameter when dealing with RNNs. LSTMs are dynamic neural networks that deviate from traditional RNNs through their use of an internal cell state, specifically its input, forget, and output gates [93] (see Figure 3.5).

In Figure 3.5, x refers to the input to the cell, c refers to the cell's internal state/memory, and h is the output. t and t-1 denote the current and previous step, respectively. The three σ nodes refer

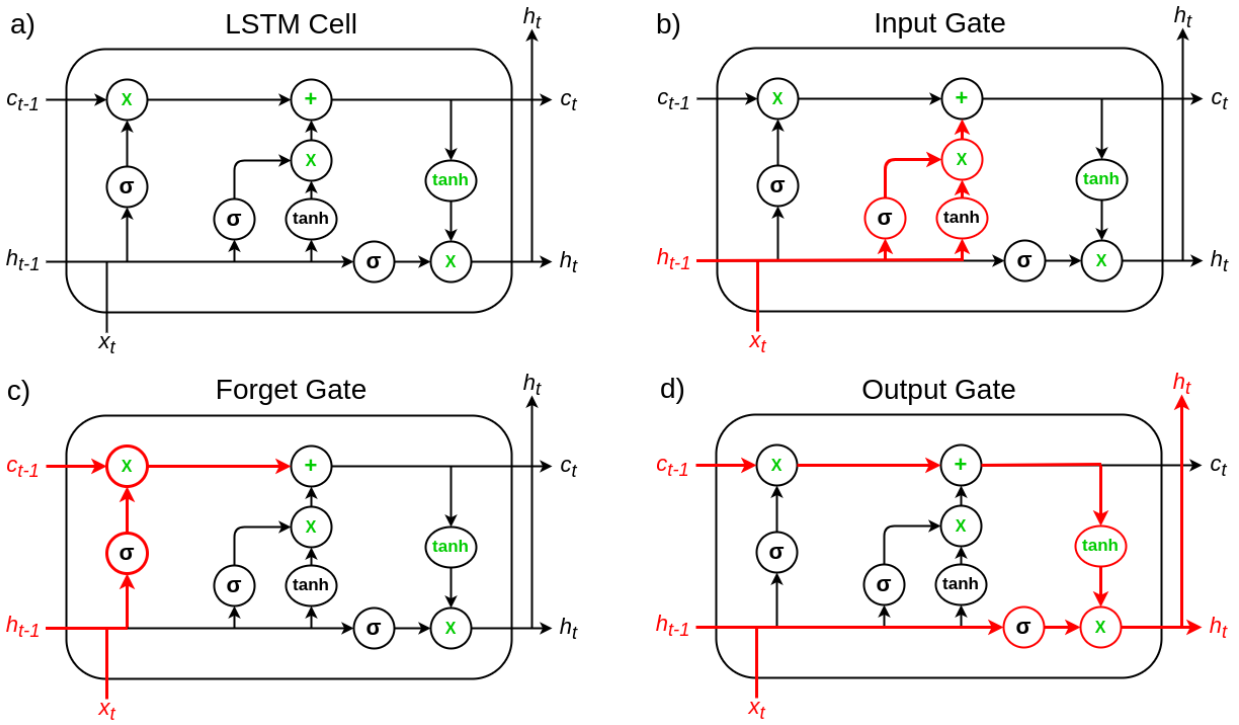

Figure 3.5: Overall construction of the LSTM cell (a) with the input gate (b), forget gate (c), and output gate (d) highlighted in red. Green is used to denote point-wise operations.

to internal layers with the sigmoid activation function, and tanh refers to either a layer with the tanh activation function or a point-wise operation, dependant on the color. The LSTM cell (a) is shown with each gate highlighted: input gate (b), forget gate (c), and output gate (d).

The internal sigmoid layers function similar to typical binary gates. If a gate is open (1), information passes through. Conversely, if a gate is closed (0), no information gets through. As sigmoid has a continuous range between zero and one (see Section 3.1.1), it is ideal to function as a gate while keeping the LSTM cell differentiable. For the input gate, the input and previous output information get passed both the the sigmoid and tanh layers. The tanh layer acts as a normal neural network layer, while the sigmoid layer determines how much out the tanh output passes through. The forget gate passes the input and previous output information through another sigmoid layer that will interact with the previous cell state. The output of the sigmoid layer will determine how much of the previous cell state will pass through.

The last gate is the output gate. This uses the input and previous output information to de-

termine how much of the output will pass through to the next layer. While only certain parts are highlighted in (d), the output gate affects information from the other two gates. While these three gates are described independently, the information passes through the cell concurrently. The internal cell state is updated, and the information passes onto the downstream layer. The internal cell state is how LSTMs can keep track of long-term information.

3.4.1 Data Preparation

In standard feed-forward neural networks, inputs and outputs simply need to have the same length about the first axis to achieve supervised training. However, recurrent neural networks require further processing. Consider the number of inputs (n_{inp}) and number of outputs (n_{out}) for a given dataset with n samples. The concatenation will result in an array of shape $n \times (n_{out} + n_{inp})$. A new hyperparameter for RNNs is the number of lag-steps (n_{LS}). This is the number of previous time steps the model will use to make a single prediction (think short-term memory for an LSTM).

The data must be stacked, so each row contains outputs and inputs for each lag-step and the current step. This orders in least-to-most recent from left-to-right. The data will now be of the shape $n \times (n_{LS} + 1)(n_{out} + n_{inp})$. The last n_{inp} columns are then dropped as they are not needed. The data can be split into training inputs and outputs where the first $n \times n_{LS}(n_{out} + n_{inp})$ columns are inputs and the last n_{out} columns are the associated output. The final step is to reshape the input data to the shape $n \times n_{LS} \times (n_{out} + n_{inp})$.

3.4.2 Training and Evaluation

With the LSTM training data prepared for the supervised learning task, a model can be trained. Typical LSTM training involves one-step prediction. This means that for each time step, the model uses the last n_{LS} sets of true inputs and outputs to make the next prediction. As it goes through an epoch sequentially, the model will not only be updating weights in the way described for feedforward neural networks, it also updates the weights associated with the three sigmoid layers and one tanh layer within each LSTM cell. Between epochs, the LSTM internal state, c, will be changing as information passes through each cell.

LSTMs also deal with a concept known as resetting the internal cell state. When reaching a

temporal discontinuity, it is important to wipe the internal memory of every LSTM cell. Otherwise, it will be using irrelevant information to make predictions. This is typically done at the end of an epoch when it reaches the end of the training time period. Resetting the state is also critical when evaluating the model on different time periods. The approach for one-step LSTM training is shown in Figure 3.6, panel (a). Note: in Figure 3.6, the term "Outputs" is used to refer to the true data while "Pred." refers to the model predictions.

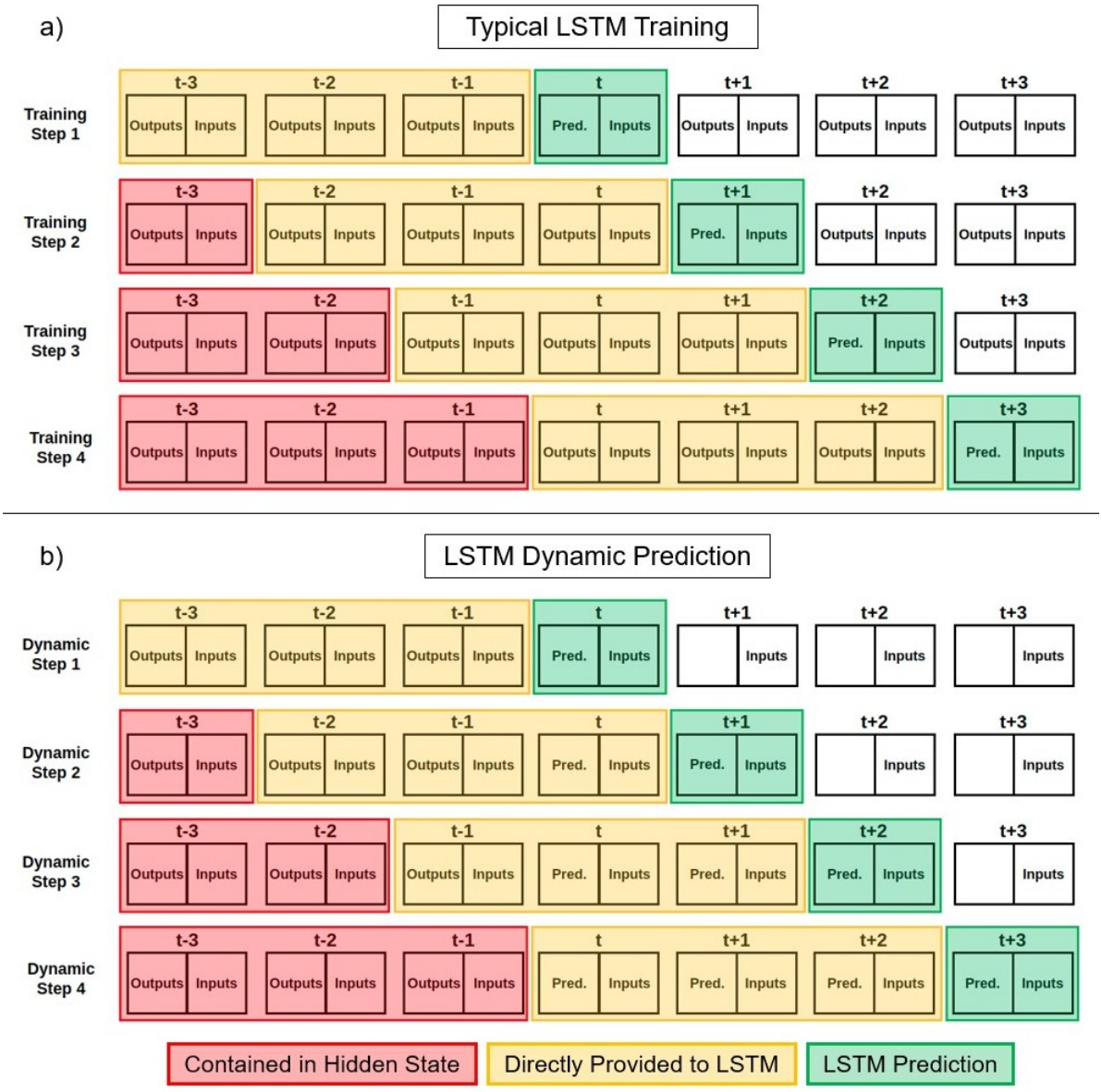

Figure 3.6: Steps for typical LSTM training (a) and steps for an iterative dynamic prediction (b) with $n_{LS} = 3$. Although the predictions and inputs for the evaluation step are highlighted in green, the model only predicts the output.

It is important to note that in the training phase, the predictions are ignored – for lack of a better term – when moving to the next time-step. The prediction is replaced with the true output for that time period. However, this is not realistic in a forecasting environment. If a forecast is required for the next 25 steps (consider $n_{LS} = 3$ from Figure 3.6), you would only have data up to the previous step *t-1*. The LSTM will make a prediction for time *t* and has to use that to make its prediction for time *t+1*. At this point, the model is taking two true values and one predicted value to make the next prediction. At *t+2*, there is only one true value and two predicted values. From *t+3* to the end of the prediction, the model will use solely the inputs and previous predictions. This is referred to as a dynamic prediction and is shown in Figure 3.6 panel (b).

3.5 Dimensionality Reduction – PCA

For a high-dimensional model (e.g. TIE-GCM and HASDM), uncertainty quantification can become infeasible in the full-state. With 12,312 and 8,760 spatial locations on the HASDM and TIE-GCM grids, respectively, a computational bottleneck forms when generating distributions for each location. To get around this, we implement a dimensionality reduction method – PCA. If the technique is implemented properly, the truncation errors can be on the order of < 3% and the reduced-state is prime for uncertainty quantification.

Principal Component Analysis is an eigendecomposition technique that determines uncorrelated linear combinations of the data that maximize variance [94, 95]. It is considered an unsupervised learning technique. PCA can be performed using the *svds* function in *MATLAB* to obtain the U, Σ, and V matrices. PCA decomposes the data and separates spatial and temporal variations such that:

$$\mathbf{x}(\mathbf{s},t) = \bar{\mathbf{x}}(\mathbf{s}) + \widetilde{\mathbf{x}}(\mathbf{s},t) \quad \text{and} \quad \widetilde{\mathbf{x}}(\mathbf{s},t) \approx \sum_{i=1}^{r} \alpha_i(t)\, U_i(\mathbf{s}) \tag{3.4}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the model output state, $\bar{\mathbf{x}}$ is the mean, $\widetilde{\mathbf{x}}$ is the deviation about the mean, r is the choice of order truncation, α_i are temporal coefficients, and U_i are orthogonal modes or basis functions. The modes are the first r columns of the left singular vector derived by performing PCA

on an ensemble of model output solutions such that:

$$\mathbf{X} = \begin{bmatrix} | & | & | & & | \\ \widetilde{\mathbf{x}}_1 & \widetilde{\mathbf{x}}_2 & \widetilde{\mathbf{x}}_3 & \dots & \widetilde{\mathbf{x}}_m \\ | & | & | & & | \end{bmatrix} \quad \text{and} \quad \mathbf{X} = U\Sigma V^T \tag{3.5}$$

In Equation 3.5, m represents the ensemble size, and the data is denoted by $\mathbf{X}$. U is the left unitary matrix, and it 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. We can extract temporal coefficients by performing matrix multiplication between Σ and V^T. Therefore, the signs of the modes and coefficients are important in the analysis phase. It is important to note that prior to centering and PCA, the density data undergoes a logarithmic transformation (log_{10}) to reduce its variance. The reader is referred to Bjornsson and Venegas [96] for more details on the distinction between PCA and Singular Value Decomposition (SVD) for use on climatic data. Figure 3.7 depicts the process of developing and using a machine-learned reduced order model based in PCA.

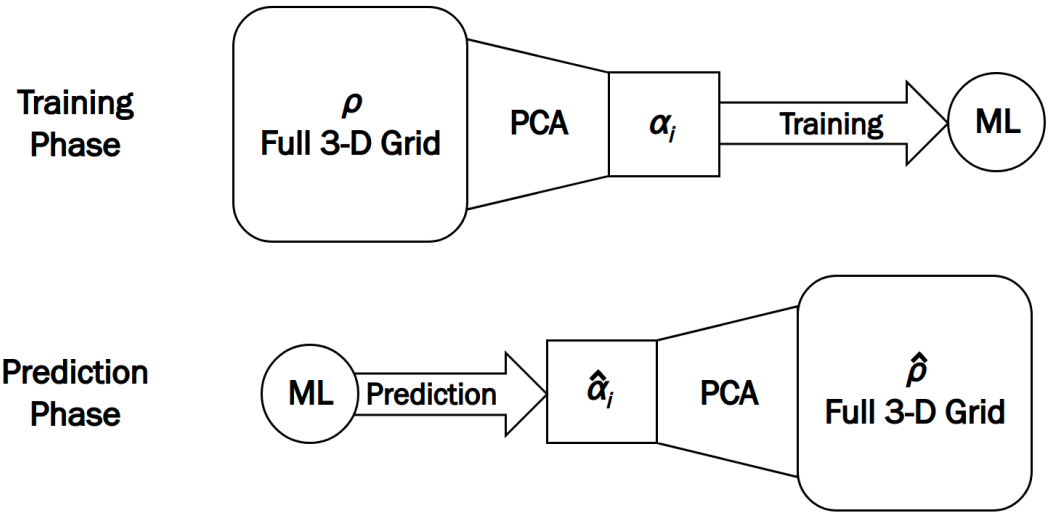

Figure 3.7: Training and prediction diagram for a PCA-based reduced order model. ρ and α denote thermospheric density and PCA coefficients, respectively.

3.6 Uncertainty Quantification

A traditional approach to regression modeling with uncertainty quantification (UQ) is to develop Gaussian process (GP) models. Gaussian process regression models are a supervised learning technique that can provide predictions with probability estimates in both regression and classification tasks [97]. A GP model essentially provides a distribution of functions that fit the data, which allows us to obtain a mean and variance for any prediction [98]. GP regression has been used recently in space weather applications in both driver forecasting [99] and empirical model calibration [100]. Some limitations for GP implementations are difficult interpretation of results for multivariate problems and computational cost with large datasets [101].

A newer approach for uncertainty quantification is Monte Carlo dropout. Dropout is a regularization tool often used in machine learning (ML) to prevent the model from overfitting to the training data [102]. 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, 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 sample is "true", 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 [103]. 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) with active dropout, 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 [104] show that MC dropout is a Bayesian approximation of a Gaussian process. Other approaches to UQ are explored in this work and will be described in the following chapter.

Chapter 4. Machine Learning for Thermospheric Mass Density

This chapter pertains to the detailed description, methodology, and analysis of the four probabilistic thermospheric mass density models developed in this work. They are used in later studies to both gain scientific insight (Chapter 5) and investigate practical impacts of uncertainty for STM (Chapter 7).

4.1 HASDM-ML

HASDM-ML is based on the PCA coefficients derived from the SET HASDM density database between 2000 and 2020 [57].

4.1.1 Methodology

The data (drivers and density) is split into training, validation, and test data using 60%, 20%, and 20%, respectively as shown in Figure 4.1. Table 4.1 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 Mehta [105].

Table 4.1: Number of time steps for different space weather conditions across the SET HASDM density database.

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

In Table 4.1, there is clear under-representation of geomagnetic storms in this vast dataset. This can cause limitations in model development, because over 98% of the dataset corresponds to $ap \leq 50$. Hierarchical modeling could be used for data of this nature, but we proceed with the

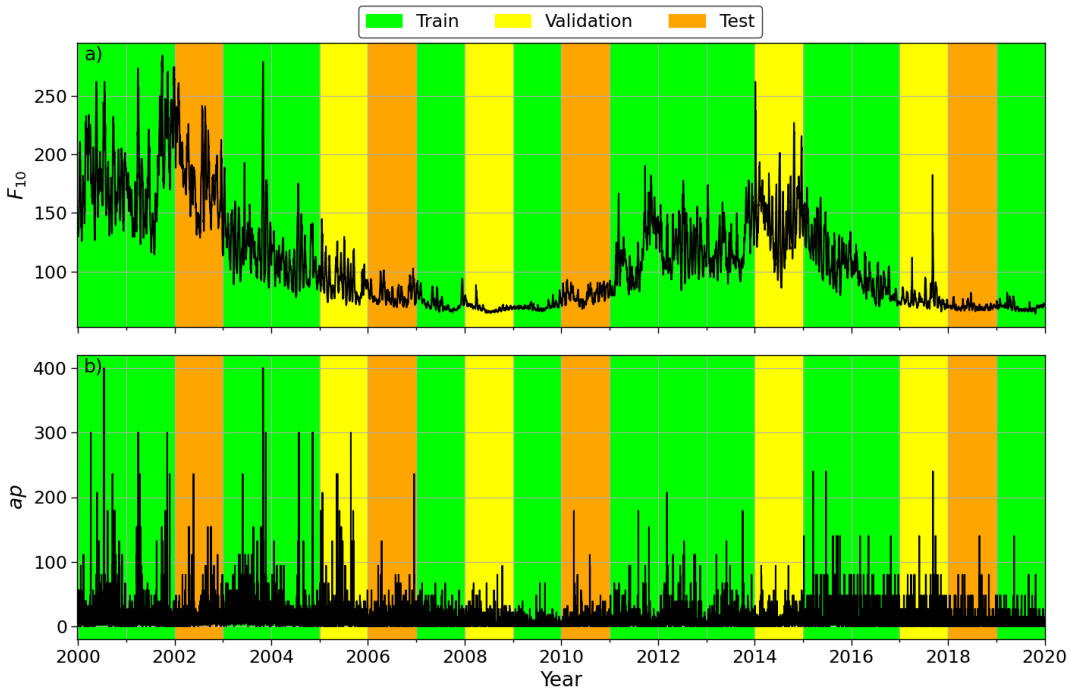

Figure 4.1: F_{10} (a) and *ap* (b) at available data points shaded to show the training, validation, and test splits.

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.

Three separate input sets are tested for the regression models, and the first two are explained in Table 4.2. The first set is the JB2008 input set, referred to as JB. However, storm and post-storm is important for the characterization of extreme events, and post-storm cooling mechanisms 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. This will be referred to as JB_H (Historical JB2008). Unlike the actual JB2008 inputs, all input sets used here contain sinusoidal transformations to the day of year (doy) and universal time (UT) inputs (shown in Equation 2.1).

In the JB_H set, the geomagnetic indices are extensive in an effort to improve storm-time and post-storm modeling. The *ap* time series is the same one used in NRLMSIS 2.0. The numerical subscript notation was previous described in Section 2.3.1.1. Early studies showed that using different time histories of *ap* and *Dst* (shown in Table 4.2) resulted in generally more calibrated

Table 4.2: List of inputs in the first 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

models (see Section 4.1.2). 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 4.2, both geomagnetic indices using the NRLMSIS 2.0 time series. This input set will be referred to as JB_{H0}.

4.1.1.1 Hyperparameter Tuning for HASDM-ML

The number of samples in the dataset is feasible for the tuner. Therefore, it is provided the full training and validation sets of 35,064 and 11,688 samples, respectively. The tuner uses 100 trials with three repeats and has the first 25 trials for the random search. A tuner is run for all three input sets and for all three loss functions tested which are described in the following section. This results in ten models for all nine input-loss combinations. The best model for each combination (based on training and validation performance) is used for the comparison in Section 4.1.2.

4.1.1.2 Uncertainty Quantification using Monte Carlo Methods

HASDM-ML had been originally developed using Monte Carlo dropout, as described in Section 3.6. 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 [106]. 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 negative logarithm of predictive density (NLPD) and continuous ranked probability score (CRPS). NLPD is based on the logarithm of the probability density

function (pdf) of the Gaussian distribution, and is shown in Equation 4.1 [107, 108].

$$NLPD(y, \mu, \sigma) = \frac{(y-\mu)^2}{2\sigma^2} + \frac{log(\sigma^2)}{2} + \frac{log(2\pi)}{2} \tag{4.1}$$

In Equation 4.1, 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 4.2 [109]. The main difference between NLPD and CRPS is that CRPS is also based on the cumulative distribution function (cdf) of the Gaussian distribution as opposed to only the pdf.

$$CRPS(y, \mu, \sigma) = \sigma \left[\frac{y-\mu}{\sigma} \text{ erf} \left(\frac{y-\mu}{\sqrt{2}\sigma} \right) + \sqrt{\frac{2}{\pi}} \exp \left(-\frac{(y-\mu)^2}{2\sigma^2} \right) - \frac{1}{\sqrt{\pi}} \right]. \tag{4.2}$$

An important aspect of using the loss functions described in Equations 4.1 and 4.2 is the preparation of the training data. The data is traditionally set up as follows. The features are set up as the number of samples (n), with n_{inp} denoting the number of inputs, resulting in the input shape ($n \times n_{inp}$). The labels are set up as the number of samples with n_{out} being the number of outputs, resulting in the output shape ($n \times n_{out}$). 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 ($n \times k \times n_{inp}$) and ($n \times k \times n_{out}$), respectively. This allows us to determine the mean and standard deviation for each sample in the batch within the loss function.

4.1.1.3 Latent Space UQ for HASDM-ML

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 Anderson et al. [110], shown as

$$\text{Calibration Error Score} = \frac{100\%}{r \cdot q} \sum_{i=1}^{r} \sum_{j=1}^{q} \left| p(\alpha_{i,j}) - p(\hat{\alpha}_{i,j}) \right| \tag{4.3}$$

In Equation 4.3, q 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 4.4,

$$p\left(\hat{\alpha}\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\hat{\alpha}_i^l < \alpha_i < \hat{\alpha}_i^u\right) \tag{4.4}$$

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 1 if the inequality is true and a 0 otherwise. To compute the bounds, we use the pair of equations given in Equation 4.5 [111, 112].

$$\hat{\alpha}_i^l = \mu - z\sigma \quad \text{and} \quad \hat{\alpha}_i^u = \mu + z\sigma, \tag{4.5}$$

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

$$z = \sqrt{2}\, \text{erf}^{-1}\left(\text{PI}\right) \tag{4.6}$$

where PI is the prediction interval of interest (e.g. 95% or 0.95). Comparing the expected ($p(\alpha)$) and observed ($p(\hat{\alpha})$) cumulative probabilities is done qualitatively by plotting the calibration curves: $p(\hat{\alpha})$ vs $p(\alpha)$. 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 4.3) 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 4.3 – 4.6, but for density, α would be replaced with ρ. **Note: this approach is used to determine calibration for all ML models in this work.**

4.1.1.4 Density UQ for HASDM-ML

While latent space calibration is important because the model is trained on the 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 and GRACE. We use the satellite **positions** for density calibration assessment between HASDM and HASDM-ML.

HASDM-ML was evaluated 1,000 times every 3 hours across the entire availability of CHAMP (2002–2010) and GRACE (2002–2011) position data listed in the measurements presented by Mehta et al. [48] and interpolated 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, the model and database density was only interpolated every 500 samples (5,000 and 2,500 seconds 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 three 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).

Geomagnetic storms are particularly difficult conditions to model accurately. Therefore, we look at four geomagnetic storms from 2002 – 2004 where HASDM-ML's reliability can be evaluated across unique events. Two of the events take place in 2002, which is outside the training dataset, while the other two are from 2003 and 2004 and are seen in training. Information on these storms can be found in Table 4.3. The model is evaluated over a 6-day period encompassing a storm then interpolated to the CHAMP locations (10 second 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 4.6 in an orbit-averaged form. We also show the orbit-averaged JB2008 predictions for comparison. In total, the six days amounts to 48 model prediction epochs which results in 51,840 interpolated densities (1,000 MC runs) from which we compute the observed cumulative probabilities. Note: this is done explicitly for the original MC dropout version of HASDM-ML

4.1.2 Results

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 4.4. The mean absolute error

Table 4.3: Information on the four storms used in the calibration analysis.

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

is computed for the model prediction (mean coefficient predictions converted to density through PCA) relative to the HASDM database.

Table 4.4: 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%

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 4.1.1, the motivation for using the time series geomagnetic indices was to improve storm-time and post-storm performance. However, Table 4.1 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 4.5 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 4.3).

The incorporation of the custom loss functions reduce the calibration error score by an order of magnitude relative to models trained with MSE, which tend to underestimate the uncertainty.

Table 4.5: Calibration error score (see Equation 4.3) 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%

The best performing loss function, in regards to calibration, is NLPD. To choose the best overall model, we focus on the test performance as that data is 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 4.4), 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 4.2.

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 4.5. There is a significant underestimation of the uncertainty at all cumulative probability levels, because the model is not trained with any terms for its variance.

The JB_H + NLPD model shown in Figure 4.2 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 opera-

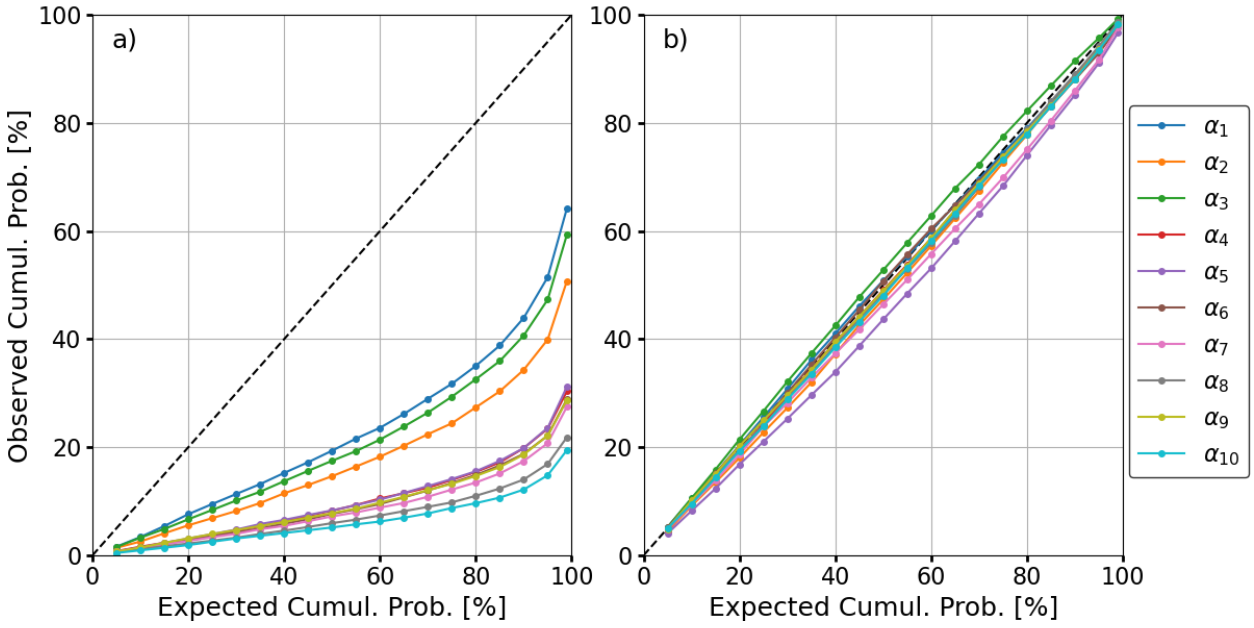

Figure 4.2: Expected vs observed cumulative probability of all 10 PCA coefficients for HASDM-ML on the test set using JB_H + MSE (a) and JB_H + NLPD (b).

tional nature, we look at the orbits of CHAMP and GRACE-A each over eight year periods with a cumulative altitude range of 300–530 km. HASDM-ML was evaluated in three-hour increments from 2002–2011, and was interpolated to the satellite positions at all epochs discussed in Section 4.1.1.4. The results for the CHAMP orbit are displayed in Figure 4.3.

Figure 4.3 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 4.4, but panels (b) and (d) show that the density uncertainty is reliable. The calibration curve is exceptional with the observed cumulative probability being within 1% of the expected value for all 20 cumulative probability levels (PIs) tested. Figure 4.4 shows the same analysis along the GRACE-A orbit.

Figure 4.4 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

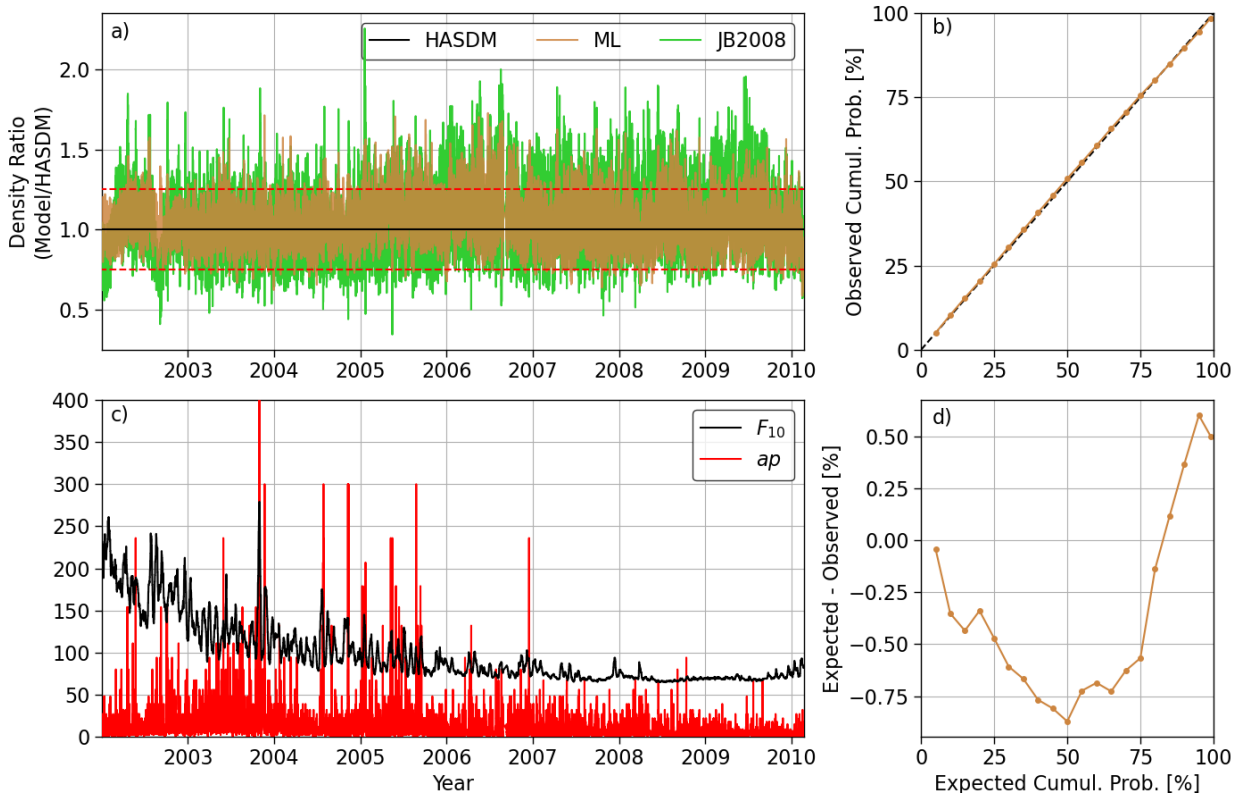

Figure 4.3: (a) shows the density ratios of HASDM-ML and JB2008 relative to HASDM, (b) shows the expected vs observed calibration curve, (c) shows F_{10} and ap for the period corresponding to (a) for reference, and (d) shows the difference between expected and observed cumulative probability corresponding to (b). Discontinuities in (a) and (c) represent data gaps. In panel (a), the red dashes lines are at ratios of 0.75 and 1.25.

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 the 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 4.5.

Figure 4.5 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 4.3 and

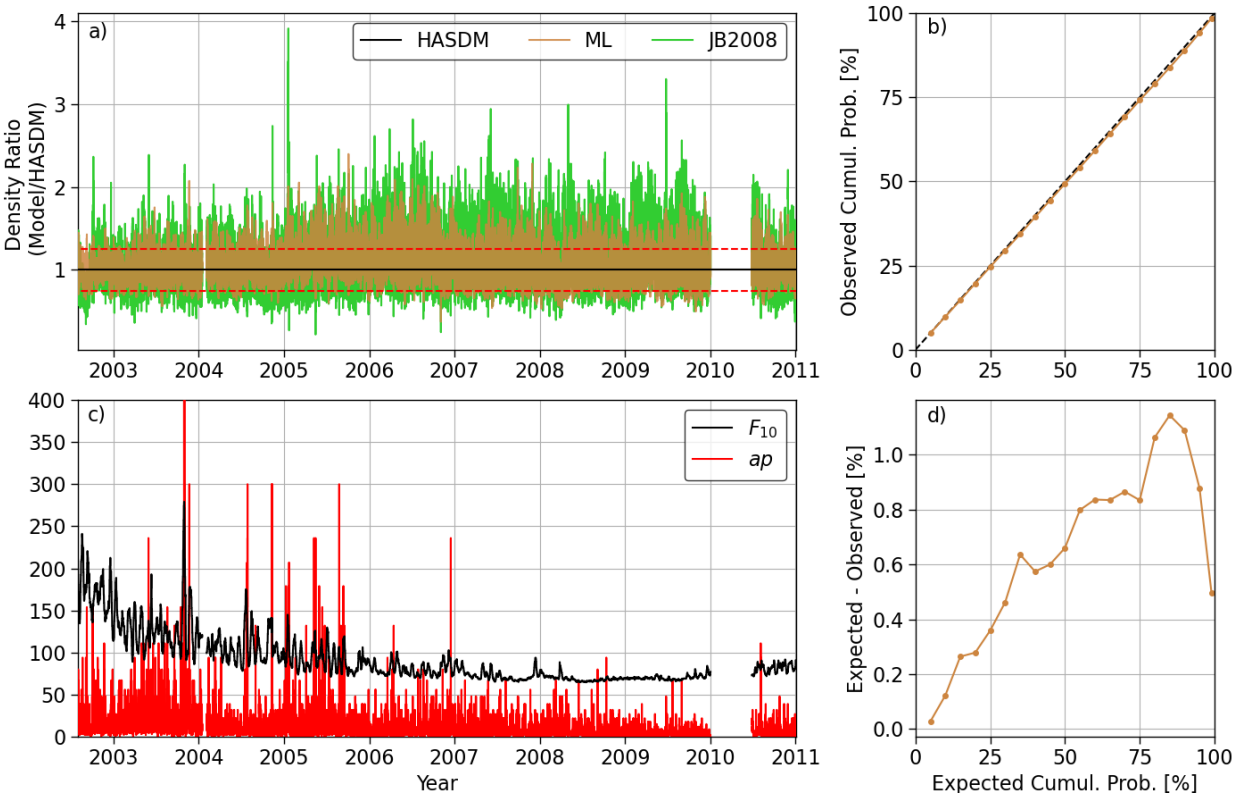

Figure 4.4: (a) shows the density ratios of HASDM-ML and JB2008 relative to HASDM, (b) shows the expected vs observed calibration curve, (c) shows F_{10} and ap for the period corresponding to (a) for reference, and (d) shows the difference between expected and observed cumulative probability corresponding to (b). Discontinuities in (a) and (c) represent data gaps. In panel (a), the red dashes lines are at ratios of 0.75 and 1.25.

4.4, 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 4.6 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 4.5). Panels (a) and (b) represent storms not con-

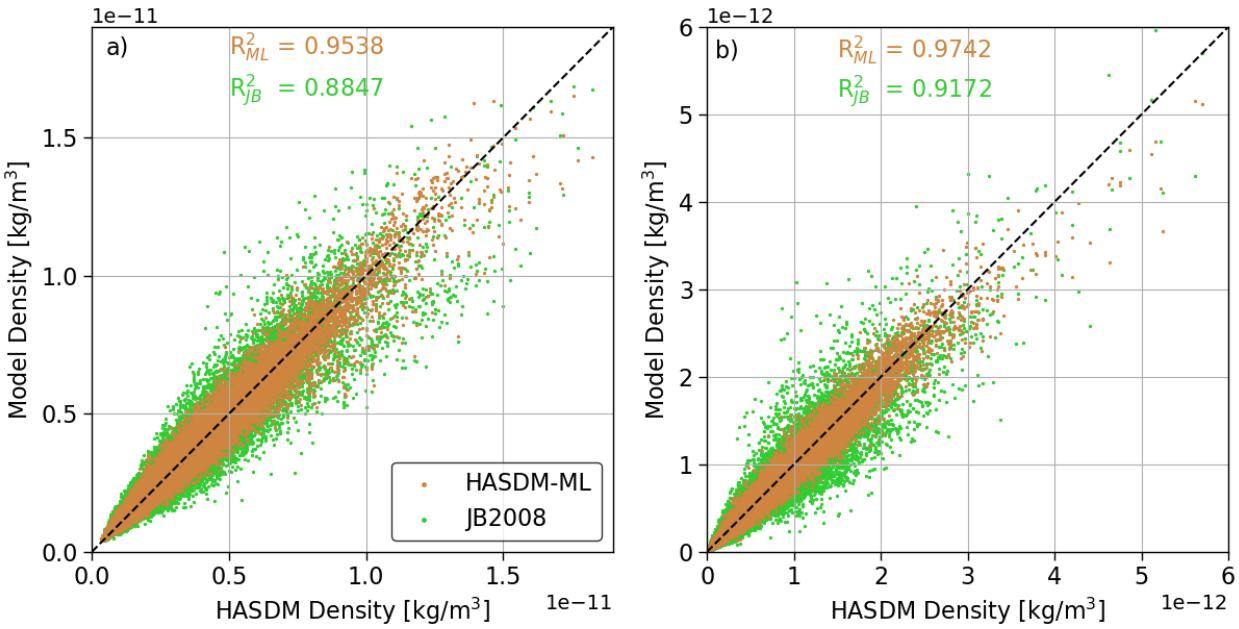

Figure 4.5: Scatter plot of model vs HASDM density along the orbits of CHAMP (a) and GRACE-A (b). Perfect prediction would fall on the diagonal black line. The coefficient of determination (R^2) is shown for both models relative to HASDM. Note: ML refers to HASDM-ML while JB refers to JB2008.

tained in the training dataset 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 model 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. Other studies compared the orbit-averaged densities of HASDM and JB2008 to CHAMP and GRACE-A during this storm and found that the low post-storm densities predicted by HASDM were similar to the density estimates from both satellites which HASDM-ML is also showing [113]. The errors for this storm are 8.46%

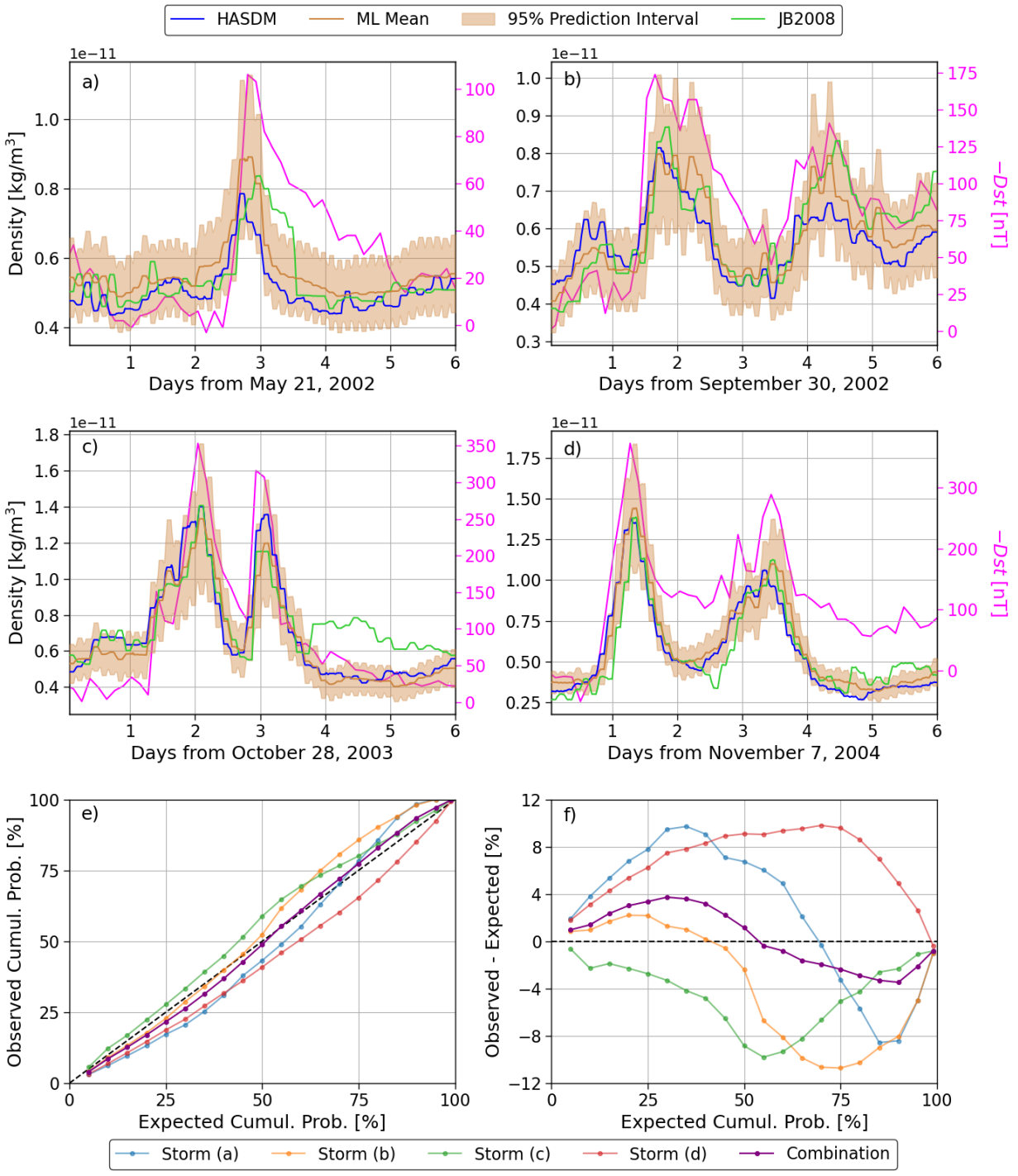

Figure 4.6: Panels (a), (b), (c), and (d) show HASDM, HASDM-ML mean, and JB2008 orbit-averaged density for CHAMP's orbit across various geomagnetic storms. The shaded region represents the 95% prediction interval for HASDM-ML, and -*Dst* is shown on the right axis in each panel. Panel (e) shows the calibration 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).

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 hours. 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 probability levels (PIs) tested, deviated from perfect calibration by more than 10.7%. The combination of all four calibration curves (averaged) is shown 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.

4.1.2.1 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 4.6 and the number of samples in each bin can be found in Table 4.1. The bottom half of the table contains the global mean absolute errors of JB2008 relative to HASDM for comparison.

The results from Table 4.6 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 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

Table 4.6: Mean absolute error across global grid for HASDM-ML and JB2008 relative to the HASDM database as a function of space weather conditions.

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

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.

4.2 Deterministic Uncertainty Quantification

While the first chosen approach to UQ was through MC dropout, another way to represent uncertainty is to directly predict the mean and standard deviation of each output. The mean square error loss function cannot be used here as there are no labels for the standard deviation. However, Nix and Weigend [114] used a neural network to directly predict the mean and variance of a toy dataset using the NLPD loss function. We implement this technique for the datasets presented. To accomplish this, we create a custom output layer with $2n_{out}$ neurons. The first n_{out} neurons represent the mean prediction and have a linear activation function. The last n_{out} neurons represent the standard deviation and use the softplus activation function. The softplus function and its derivative – the sigmoid function – are shown in Equation 4.7.

$$f(x) = ln(1 + e^x) \qquad f'(x) = \frac{e^x}{1 + e^x} \tag{4.7}$$

The desired qualities of the standard deviation output are: (1) always positive and (2) having no upper bound. The initial choice was the absolute value function. However, the resulting models had erratic loss values, and it was difficult to obtain a good model. The softplus function is (1) always positive, (2) has no upper bound, (3) is monotonically increasing, and (4) is differentiable across all inputs. This resulted in stable training losses and better models.

This technique will be compared to MC dropout for its validity in terms of performance for both HASDM-ML and later CHAMP-ML (see Section 4.3). It will also be vital as a potential approach for TIE-GCM ROPE (Section 4.5), because MC dropout with NLPD would be too computationally expensive.

4.2.1 Direct Probability Prediction Toy Example

As previously mentioned, the uncertainty distribution can be directly predicted by the model. To visualize the way this work with the NLPD loss function, we train basic models for two toy problems. Each problem is a function, $y = f(x)$, with additive Gaussian noise having zero-mean and a functional form to the standard deviation. These functions are displayed in Table 4.7. The results for Problem 1 is shown in Figure 4.7

Table 4.7: Functions for the two toy problems with the right column being the functional form of the Gaussian noise.

	Function	$\boldsymbol{\sigma}$
Problem 1	$0.3x + cos(0.5x) - 4 + \mathcal{N}(0, \sigma)$	$0.5\frac{e^{sin(0.2x)}}{1+e^{sin(0.8x)}}$
Problem 2	$sin\left(2x + cos(3x)\right) + \mathcal{N}(0, \sigma)$	$0.05sin(0.2x) + 0.025$

Figure 4.7 shows that the model is able to adequately predict the function and is able to predict the overall probability distribution. The important aspect of the figure is panel (d): the model is able to predict the standard deviation without a label. Meanwhile, this is fairly trivial data. Figure 4.8 shows the predictions and calibration curve for the more complex Problem 2.

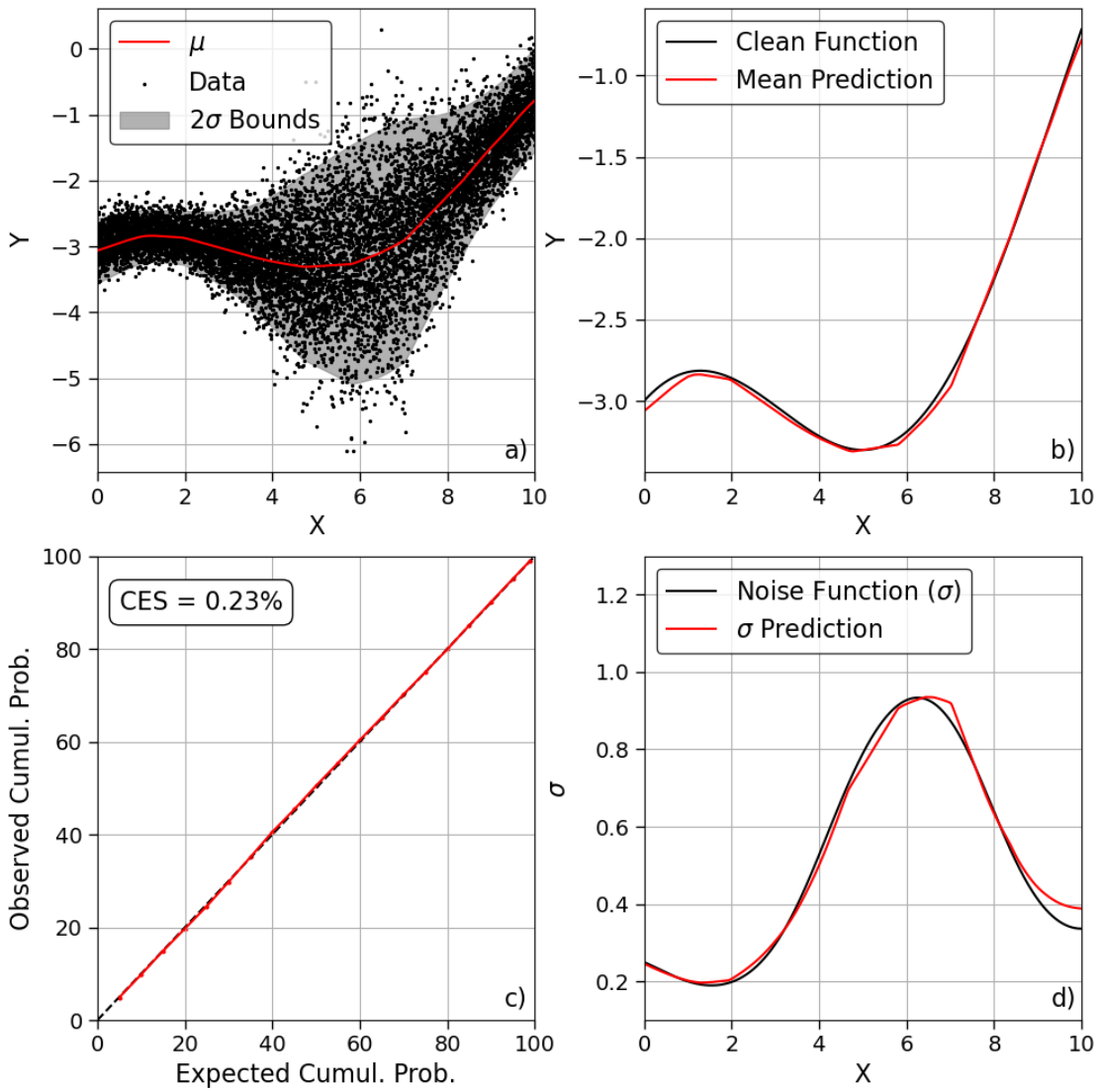

Figure 4.7: Mean prediction with 2σ bounds plotted on data (a), clean function plotted with mean prediction (b), calibration curve (c), and predicted standard deviation on true standard deviation function (d) for Problem 1.

For the more complex data, the model is not as accurate over all x. When $x \in [4, 6]$, the model can accurately predict the mean and standard deviation. When $x > 6$, the standard deviation prediction no longer represents uncertainty in the data but the model's uncertainty in its prediction. This is also the case for $x < 4$, but the predictions follow the general trend and σ is closer to the truth. For this portion of panel (b), the mean prediction deviates from the true mean of the data and the standard deviation in panel (d) consequently increases. Panel (c) shows that the model is

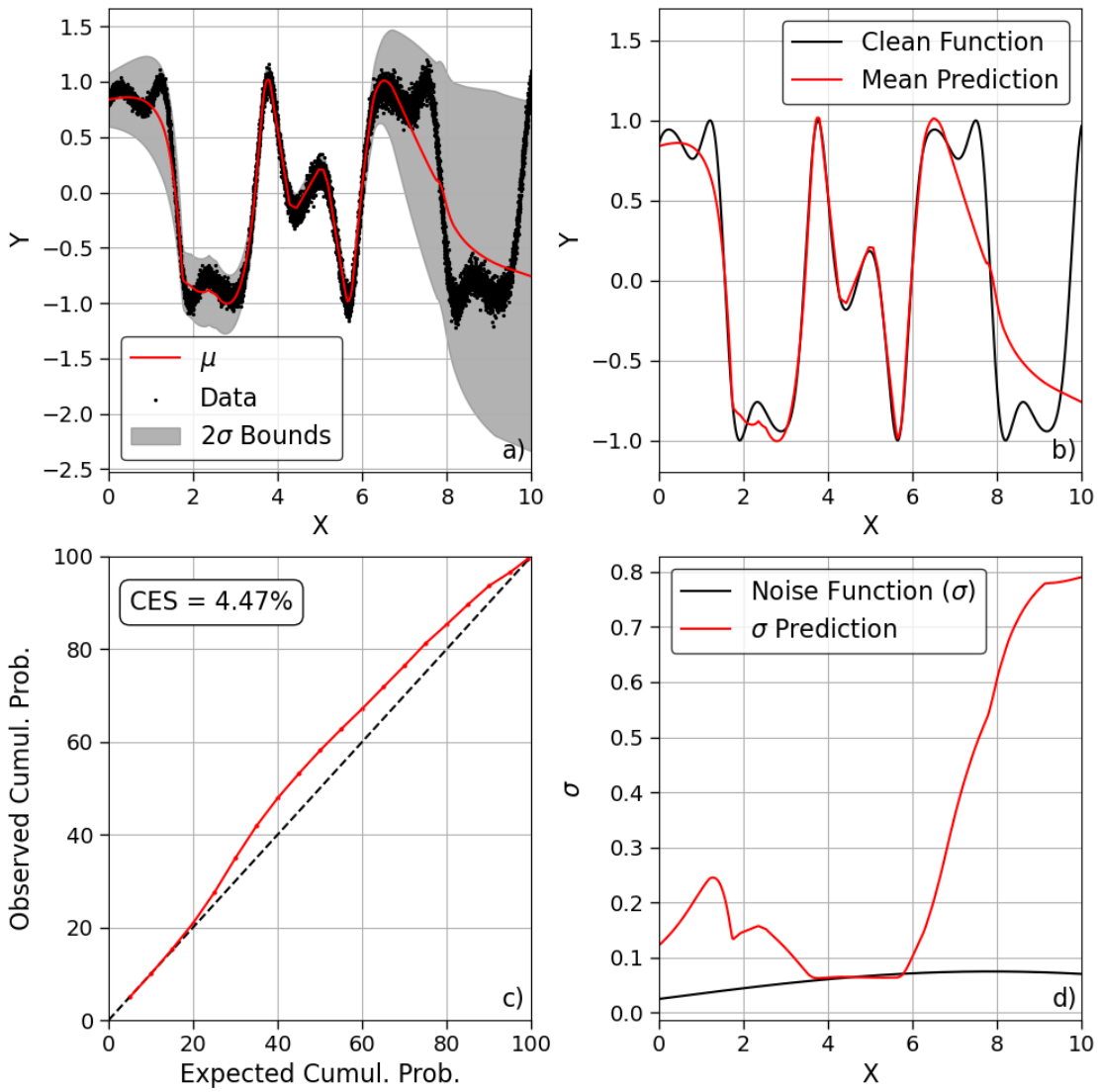

Figure 4.8: Mean prediction with 2σ bounds plotted on data (a), clean function plotted with mean prediction (b), calibration curve (c), and predicted standard deviation on true standard deviation function (d) for Problem 2.

still well-calibrated and representing both uncertainty in the data and uncertainty in the model's predictions.

The NLPD loss function does not ensure model calibration. However, we show that it can be used – if properly tested – in model development to represent uncertainty in the data and uncertainty in the model's predictions. Note: these models were trained on the entire dataset, and this is purely for demonstration. The thermospheric density models are developed with separate

validation and independent test sets.

4.2.2 Direct Mean-Standard Deviation Prediction for HASDM-ML

Using the best tuner models for MC dropout and direct probability distribution prediction, we assess the error and calibration statistics. Table 4.8 shows the mean absolute error and calibration error score for both techniques across the training, validation, and test sets.

Table 4.8: HASDM modeling results using MC dropout and direct probability prediction. Error refers to mean absolute error, and calibration is computed using Equation 4.3.

Metric	Set	MC Dropout	Direct Probability
Error	**Training**	9.07%	**8.55%**
	Validation	10.69%	**9.91%**
	Test	10.69%	**10.60%**
Calibration	**Training**	3.06%	**1.74%**
	Validation	2.51%	**2.45%**
	Test	**1.76%**	2.81%

It is evident that the performance using both methods is very similar. Across all three sets, the mean absolute error and calibration error score do not deviate by more than 0.8% and 1.4% respectively. The MC dropout model has better performance on the independent test set in terms of calibration. This is a desired quality as the test data is not used for model development in any way. As the calibration error scores are composites of the scores for each output, the calibration curves are shown in Figure 4.9 for a qualitative assessment.

Both techniques lead to slightly overestimated uncertainties on the training set for multiple outputs. Meanwhile, the remaining outputs are almost perfectly calibrated. On the validation set, each model has outputs with overestimated and underestimated uncertainties. Again, most of the outputs are very well-calibrated which is affirmed by the calibration error scores. For the test set, the direct probability prediction model tends to marginally underestimate the uncertainty while the MC dropout model provides reliable uncertainty estimates on virtually all model outputs. Table 4.9 shows the mean absolute error for both models across an array of solar and geomagnetic conditions.

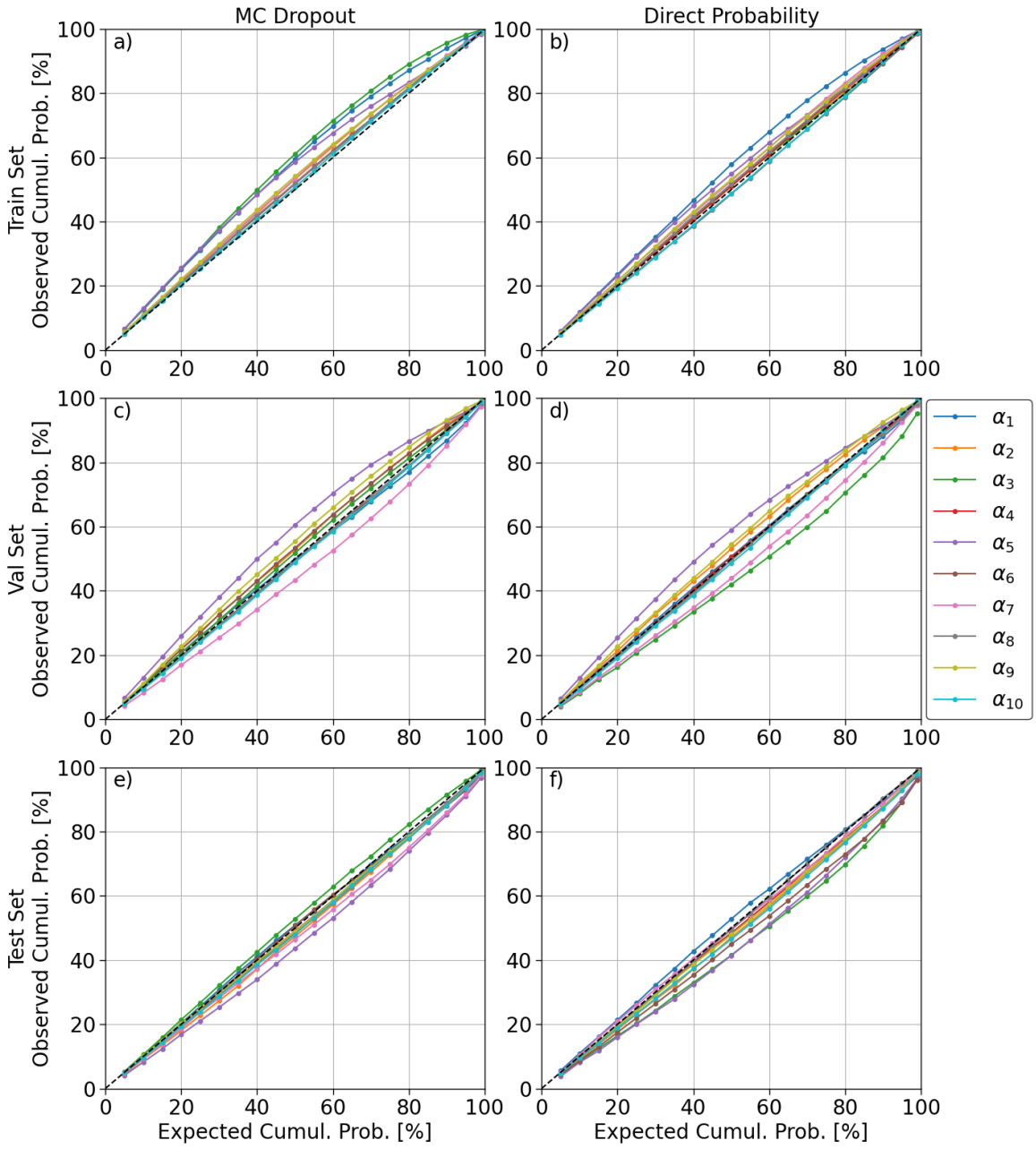

Figure 4.9: The left and right columns show the MC dropout and direct probability calibration curves, respectively. The top, middle, and bottom rows are the calibration curves for the training, validation, and test sets, respectively.

The entire dataset is used for this analysis as there are not enough samples in each bin using only the test set.

These errors tend to reiterate the results from Table 4.8. The direct probability model was more

Table 4.9: Mean absolute error across global grid for HASDM-ML as a function of space weather conditions.

MC Dropout					
	$\mathbf{F_{10} \leq 75}$	$\mathbf{75 < F_{10} \leq 150}$	$\mathbf{150 < F_{10} \leq 190}$	$\mathbf{F_{10} > 190}$	**All F_{10}**
$\mathbf{ap \leq 10}$	8.96%	9.78%	9.97%	9.14%	9.50%
$\mathbf{10 < ap \leq 50}$	9.76%	10.05%	10.87%	9.90%	10.09%
$\mathbf{ap > 50}$	15.35%	12.86%	13.23%	12.55%	13.01%
All ap	9.12%	9.92%	10.36%	9.55%	9.71%
Direct Probability					
	$\mathbf{F_{10} \leq 75}$	$\mathbf{75 < F_{10} \leq 150}$	$\mathbf{150 < F_{10} \leq 190}$	$\mathbf{F_{10} > 190}$	**All F_{10}**
$\mathbf{ap \leq 10}$	8.64%	9.33%	9.35%	9.11%	9.10%
$\mathbf{10 < ap \leq 50}$	9.18%	9.51%	9.69%	9.64%	9.48%
$\mathbf{ap > 50}$	11.14%	11.23%	11.34%	10.30%	11.11%
All ap	8.74%	9.42%	9.52%	9.34%	9.23%

accurate on all three sets, and Table 4.9 shows that it is also more accurate across all 20 conditions considered. For a majority of the conditions, the difference is small (< 1%). However, the high *ap* conditions show that the direct probability model makes considerable improvements. These error reduction from MC dropout range from 1.6 – 4.1%.

To further assess the uncertainty capabilities of the models, we attempt to visualize the calibration in the full-state (global density grids) to identify any spatial dependence in the reliability of the uncertainty estimates. First, the models are evaluated on the entire test set and the density mean and standard deviations are extracted. Using these statistics, the observed cumulative probability with a 90% prediction interval is computed for each spatial location. The resulting $24 \times 19 \times 27$ array is used to determine how well calibrated the model is on independent data as a function of location. We show seven maps for each model (200, 300, ... , 800 km) in Figure 4.10. Even though HASDM has a lateral spatial resolution of 24 longitude and 19 latitude segments, we interpolate the results to the polyhedral grid used in the EXTEMPLAR model for visualization purposes. This is done in the remainder of this work.

For reference, perfect calibration in Figure 4.10 would be uniform green maps at all altitudes. This would convey that with a 90% prediction interval, the model's predictions/uncertainty esti-

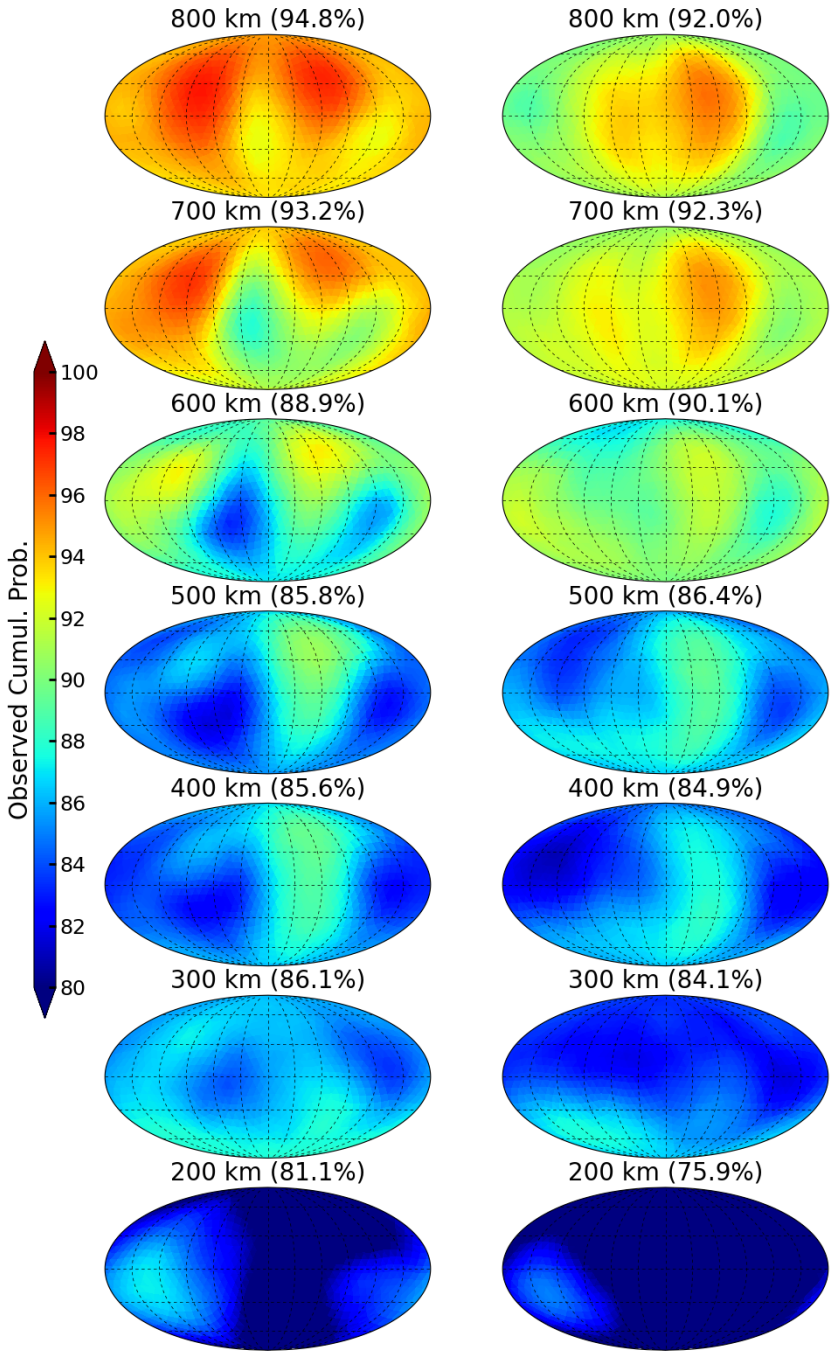

Figure 4.10: Observed cumulative probability maps for a 90% prediction interval using the MC dropout (left) and direct probability (right) models. The average observed cumulative probability is shown for each altitude in parenthesis.

mates contain 90% of true samples at all locations. While this is not the case, the results are still insightful. At 200 km, both models are underestimating the uncertainty by 10 – 15%. This could

be a result of the relative variability as a function of altitude in the SET HASDM density database. The general trend of relative variability is that it increases with altitude, so the models may underpredict the standard deviation at low altitudes as a result, which indicates that the model has a false sense of confidence in that region. Both models have an average cumulative probability within 5% of the expected value at most of the altitudes shown in Figure 4.10 with the best results at 600 km. At 700 and 800 km, both models begin to overestimate uncertainty, likely because they have the lowest confidence at those altitudes. An interesting outcome of this study is the lateral variability of the cumulative probability between the models. The MC dropout model (left) has more lateral variability, meaning the cumulative probability changes more as a function of longitude and latitude.

4.3 CHAMP-ML

While creating a surrogate model for HASDM is highly desired due to its state-of-the-art capabilities, limitations arise from its relationship with JB2008. Empirical models rely on parametric equations that can limit their ability to model some physical phenomena. A way to bypass this would be to create an entirely data-driven model. To study this, a ML model is developed directly from the CHAMP density estimates of Mehta et al. [48] after undergoing a logarithmic transformation (log_{10}).

The CHAMP dataset is significantly larger than HASDM with over 25 million total samples. Unlike the HASDM dataset, location is now an input. CHAMP only covers the local solar time domain once every three months due to its near-polar orbit. The dataset also does not span an entire solar cycle. Splitting an in-situ dataset like this using long segments – on the order of months or years – for training, validation, and testing can result in a model that is not well-generalized across the sets. Non-ML models like EXTEMPLAR and NRLMSISE-00 will also have varying error statistics across the periods as well since the conditions can be very different [115] Therefore, a different approach to data splitting was implemented. The first eight weeks are used for training (483,840 samples), then the following week is used for validation (60,480 samples), and the next week is used for the test set (60,480 samples). This scheme is repeated through the entire dataset,

resulting in similar input and output distributions while keeping temporally disjoint sets as there are two weeks or 120,960 samples between the training segments. For the tuner, 1 million random samples are chosen from the training data and 500,000 random samples are chosen from the validation data. Once the tuner is complete, the best models are retrained on the full training set and evaluated on the other two sets.

An early version of CHAMP-ML was developed using both MC dropout and direct probability prediction to compare the techniques for two unique datasets (HASDM and CHAMP). These results will be shown for comparison purposes, but a more thorough version of CHAMP-ML is also developed using specifically direct probability prediction (called CHAMP-ML-v2). The inputs for both models are shown in Table 4.10

Table 4.10: List of inputs for both versions of CHAMP-ML. *LAT* and *ALT* refer to the latitude and altitude, respectively.

CHAMP-ML			**CHAMP-ML-v2**		
Solar	**Geomagnetic**	**Spatial/Temporal**	**Solar**	**Geomagnetic**	**Spatial/Temporal**
F_{10}, S_{10}, M_{10}, Y_{10}, F_{81c}, S_{81c}, M_{81c}, Y_{81c}	*SYM-H*, S_N, S_S	LST_1, LST_2, LAT, ALT, t_1, t_2, t_3, t_4	F_{10}, S_{10}, M_{10}, Y_{10}, F_{81c}, S_{81c}, M_{81c}, Y_{81c}	S_N, S_S, *SYM-H*, *SYM-H*$_{0\text{-}3}$, *SYM-H*$_{3\text{-}6}$, *SYM-H*$_{6\text{-}9}$, *SYM-H*$_{9\text{-}12}$, *SYM-H*$_{12\text{-}33}$, *SYM-H*$_{33\text{-}57}$	LST_1, LST_2, LAT, ALT, t_1, t_2, t_3, t_4

4.3.1 Results using Both Techniques

After running tuners for both uncertainty techniques, the best models were trained on the entire training set. The models were chosen based on the lowest prediction error and best calibration scores on the validation set. Table 4.11 shows the mean absolute error and calibration error scores on the three sets.

Both models are well-generalized in terms of prediction accuracy. The range in error between sets for the MC dropout and direct probability model is 0.54% and 0.23%, respectively. Both models have higher calibration error scores on the training set but have similar scores on the val-

Table 4.11: CHAMP modeling results using MC dropout and direct probability prediction. Error refers to mean absolute error, and calibration is computed using Equation 4.3.

Metric	Set	MC Dropout	Direct Probability
Error	**Training**	13.13%	**12.59%**
	Validation	13.67%	**12.82%**
	Test	13.14%	**12.62%**
Calibration	**Training**	**3.93%**	5.84%
	Validation	0.64%	**0.25%**
	Test	**0.22%**	0.37%

idation and test sets. The two techniques provide similar results with the only notable difference is the 1.91% higher calibration error score for the direct probability model on the training set. The calibration curves for both models are shown in Figure 4.11.

Both models are well-calibrated on all three sets. There is a tendency for both models to slightly overestimate uncertainty on the training set which is more evident for the MC dropout model. The differences between the calibration curves and the perfectly calibrated reference line (in black) is shown in panels (c) and (d). Panel (d) highlights the overestimation of uncertainty for the direct probability model on the training set. However, it never deviates by more than 9%. Both models tend to underestimate uncertainty on the validation and test set for the larger prediction intervals. Again, the deviation from perfect calibration is no more than 2% for any PI. Due to the intrinsic difference between the datasets that the CHAMP and HASDM models are developed from, the proceeding analyses will be different than those in Section 4.2.2.

4.3.2 Global Modeling with Local Measurements

The CHAMP models were developed with in-situ measurements, but we hypothesize that it should be able to learn the functional relationship of the combined inputs. Therefore, the model should be able to provide global outputs at any point in time. As a qualitative assessment, we show global maps at 400 km for the winter and summer solstices in Figure 4.12 using the direct probability model. All proceeding global analyses will be performed using this model. For this test, the solar drivers are all set to 120 sfu, *SYM-H* is set to 0 nT, both Poynting flux totals are set

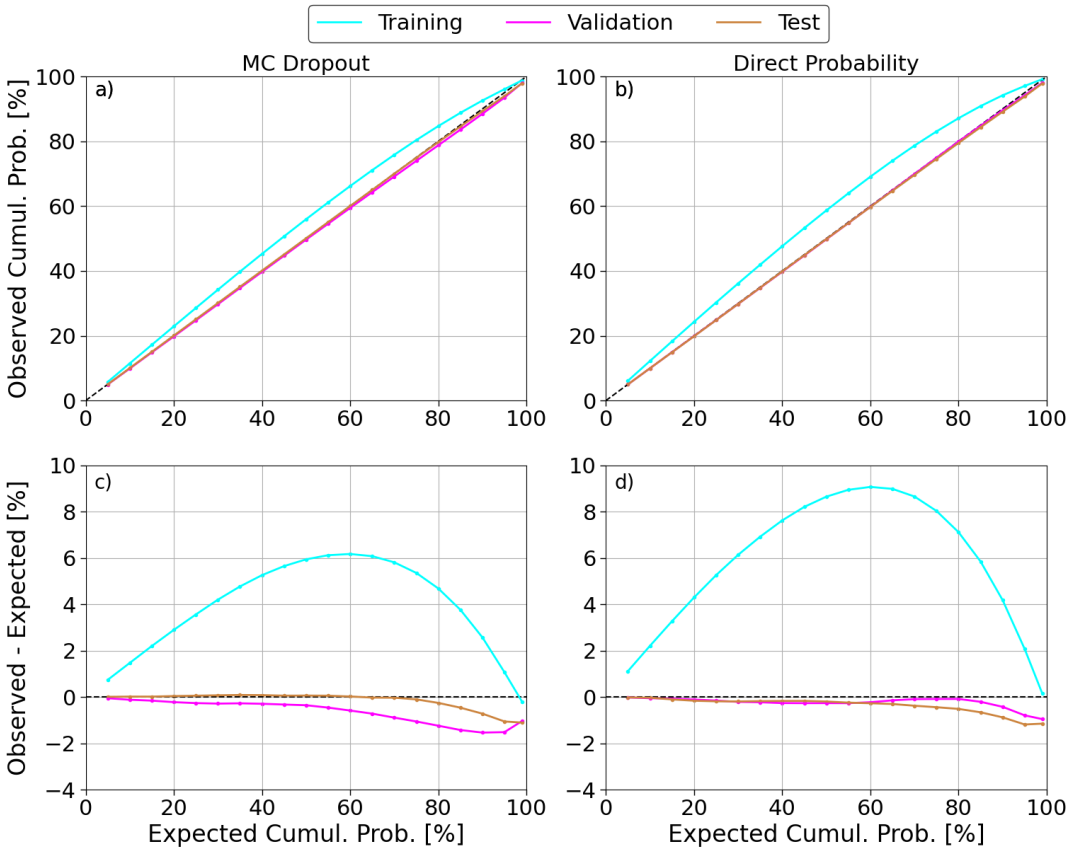

Figure 4.11: Calibration curves for the training, validation, and test sets using MC dropout (a) and direct probability prediction (b). Panels (c) and (d) show the difference between the observed and expected cumulative probability using MC dropout and direct probability prediction, respectively.

to 27GW, and the time is set to 00:00 UT.

The diurnal structure is present in both panels with the peak density being in the southern hemisphere during the winter solstice and in the northern hemisphere during the summer solstice. This shows the model's understanding on annual trends (Earth's tilt). The general density level is higher during the winter solstice, but the relative variation between day and night are very similar. This is reaffirmed by the exospheric temperature distribution shown by Weimer et al. [69] during the solstices.

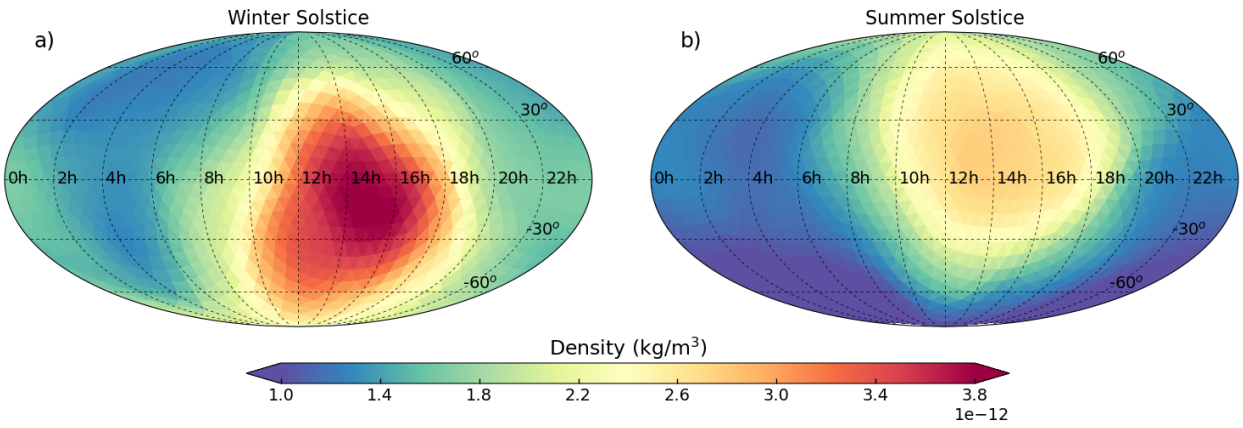

Figure 4.12: Global density map with moderate solar activity, low geomagnetic activity, the altitude fixed to 400 km, and the time of day being 00:00 UT for the winter solstice (a) and the summer solstice (b).

4.3.3 Investigating the Uncertainty

Next, we look at the uncertainty levels for eight unique conditions of activity and time. These are all displayed in Table 4.12. Using these space weather and temporal inputs, the CHAMP model is evaluated at all 1,620 polyhedral grid locations from 300 to 450 km in 1 km increments. The metric we use here is a normalized measure of model uncertainty: $100 \cdot \sigma/\mu$, essentially providing the 1-σ uncertainty as a percentage of the mean prediction. The resulting maps are averaged across each altitude to evaluate the model's uncertainty for each condition as a function of altitude. Three aspects of model drivers are investigated: solar activity, geomagnetic activity, and temporal dependence. In Table 4.12, there are three solar activity levels, with all other drivers kept constant. There are also three geomagnetic cases: low and high geomagnetic activity with moderate solar activity, and high geomagnetic activity with high solar activity. We only look at two daily cases – 00:00 and 12:00 UT. We also look at the fall equinox, summer solstice, and winter solstice with moderate solar and low geomagnetic activity. The resulting altitude profiles are shown in Figure 4.13.

Panel (a) in Figure 4.13 shows that the CHAMP model has low uncertainty in its lower altitude predictions for solar minimum (or low solar activity) which drastically increases with altitude.

Table 4.12: CHAMP model inputs to study various conditions as a function of altitude. * Solar 2 is also considered Geo 1, UT 1, and doy 1.

	Solar Drivers	Geomagnetic Drivers		Temporal Drivers	
Condition Name	**FMSY**	**SYM-H**	**$S_N = S_S$**	**UT**	**doy**
Solar 1	75	0	27	0	262
Solar 2*	120	0	27	0	262
Solar 3	190	0	27	0	262
Geo 2	120	-75	128	0	262
Geo 3	190	-75	128	0	262
UT 2	120	0	27	12	262
doy 2	120	0	27	0	172
doy 3	120	0	27	0	355

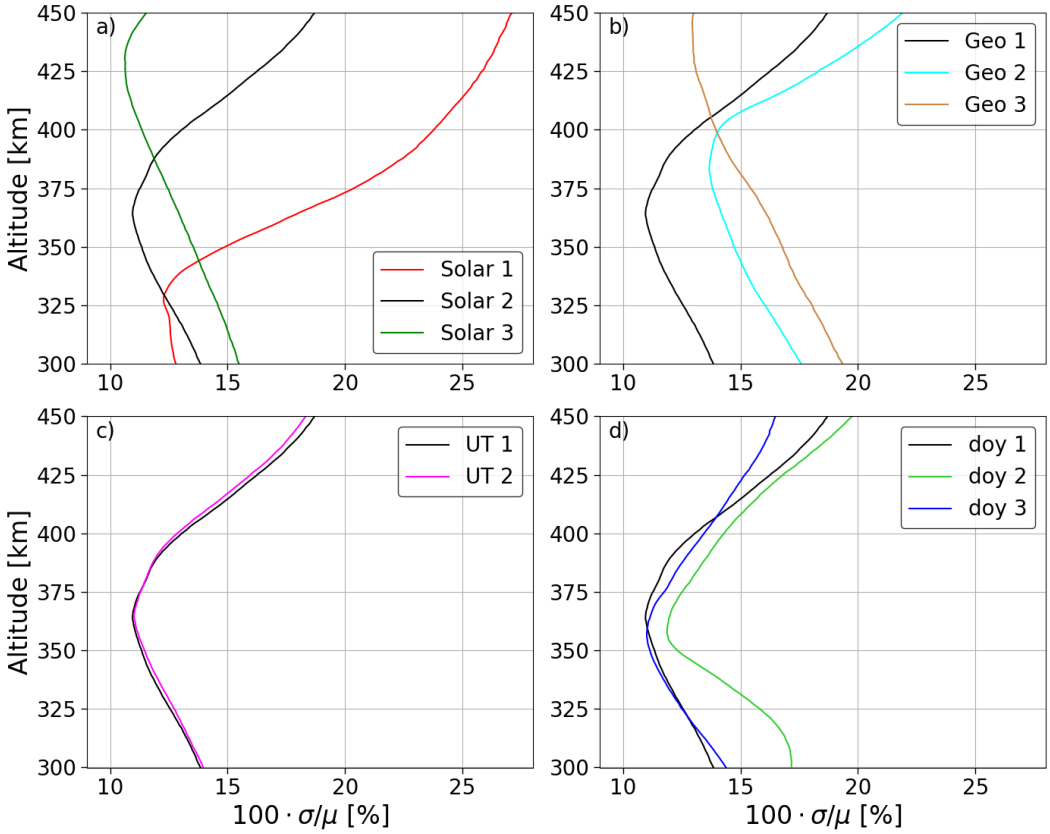

Figure 4.13: Normalized uncertainty variations as a function of altitude for solar (a), geomagnetic (b), daily (c), and annual (d) cases. The drivers for each curve can be found in Table 4.12.

The opposite can be said for solar maximum. The moderate solar activity case results in lowest uncertainties between 350 and 375 km and higher uncertainties above and below that range. This is all a result of CHAMP's altitude from 2002-2010. It started around 460 km during solar maximum and ended at 300 km during solar minimum. Therefore, the model has confident predictions in the altitude range the satellite was located during the various phases of the solar cycle. If there was additional data from satellites at different altitudes over a longer time period, the model would likely be more confident over a larger altitude range.

In panel (b), we see the same general trends for Geo 1 and Geo 2, because they are evaluated using moderate solar activity. However, it is evident that the increase in geomagnetic activity results in up to 5% more uncertainty. The Geo 3 case is similar to Solar 3 (high solar activity) but again has increased uncertainty due to the storm conditions it represents. Panel (c) indicates that there is a low impact from universal time on the model uncertainty. In Panel (d), the black line indicates the fall equinox which is similar to the winter solstice. The Winter solstice uncertainties deviate from the equinox uncertainties at the highest altitude range. While the overall shape remains consistent, there are highest uncertainties for the summer solstice at all altitudes. The overall takeaway form Figure 4.13 is that the shape of the model uncertainty altitude profile is most strongly effected by the solar activity level while the day of year and geomagnetic activity tend to uniformly increase or decrease uncertainty. These profiles would all likely be impacted if the model was developed using additional satellite data.

4.3.4 Evaluation Time Comparison

This section aims to provide an equal comparison of the two methods in terms of computational complexity. To do so, each CHAMP model is evaluated on either 8,640 samples (one week) or 86,400 samples (ten weeks). For the direct probability prediction model, it sees each input once and provides the mean and standard deviation. These are used to sample a Gaussian distribution 1,000 times to get probabilistic predictions for density over the given window.

For MC dropout, we cannot pass one week of inputs to the model stacked 1,000 times (as is done for HASDM). There is not enough memory on an NVIDIA GeForce RTX 2080 Ti graphics

processing unit (GPU) – 11 GB – to perform this evaluation. Therefore, we pass the 100 repeated inputs in 10 chunks to obtain the 1,000 predictions. When evaluating over ten weeks, we must reduce to 10 repeated inputs in 100 chunks. In Table 4.13, we show the evaluation times on both GPU and CPU for both methods over the two durations. Note: when running MC dropout on CPU, we use 100 repeated inputs for both durations. The batch size for all predictions is 2^{17} or 131,072. The size of the MC dropout and direct probability models are 233.3 kB and 21.9 MB, respectively.

Table 4.13: Run time to obtain 1,000 probabilistic predictions from each model using GPU and CPU in seconds.

Method	Samples	GPU Run Time	CPU Run Time
MC Dropout	8,640	2.11	13.65
	86,400	18.29	127.79
Direct Probability	8,640	0.58	0.52
	86,400	3.93	3.93

The run times are unique to these specific models. The size of the models plays a role in run time, and the size of these models are a result of the tuner. The MC dropout model is approximately 100 times smaller, but the increase in required model prediction calls results in the significantly longer run times. The direct probability method, for this particular problem, is anywhere from 3 to 30 times faster depending on the number of samples and whether the GPU or CPU is being used.

4.4 MSIS-UQ

CHAMP-ML is particularly important due to its truly data-driven nature with no user-defined basis functions to skew results. However, the limited altitude range of the dataset hinders its use throughout the thermosphere as STM is not limited to the 300–460 km altitude range of the CHAMP dataset. The EXTEMPLAR model [69, 70] uses a similar dataset to provide more accurate exospheric temperatures to an MSIS model. We aim to build on this work while leveraging ML to introduce nonlinearity and UQ capabilities.

4.4.1 Methodology

MSIS-UQ is a machine-learned exospheric temperature model based on the temperature estimates described in Section 2.3.2.1. This model is similar to EXTEMPLAR-ML [115] but differs by: 1) using true locations (no grid) for training, 2) using the newest MSIS model, and 3) providing uncertainty estimates. To accompany the high temporal cadence of the measurements, we expand from the input set of NRLMSIS 2.0. To account for solar activity, the model receives F_{10}, S_{10}, M_{10}, and Y_{10}, all accounting for different forms of solar emissions that affect different regions of the thermosphere . The ML model also uses inputs from EXTEMPLAR, particularly S_N, S_S, ΔT. It also uses the *SYM-H* time series from CHAMP-ML-v2 along with the same spatial and temporal inputs. The exception to this is altitude as exospheric temperature is independent of satellite altitude.

4.4.1.1 Data Preparation

The 81 million samples – inputs and $log_{10}(T_\infty)$ – are split into training, validation, and tests sets to achieve an 80%–10%–10% distribution. The dataset is split the same way as CHAMP-ML. Again, there is a significant number of samples within each segment providing temporally disjoint segments throughout the 17 year time-span of the dataset. Since each satellite has a different cadence and there are different numbers of satellites providing measurements at a given time, the number of samples in each segment varies. In the training, validation, and test sets, the number of samples varies from 22,454–1,450,380, 12,990–181,423, and 12,888–181,422, respectively. This means that there are between 25,878 and 362,845 samples separating training segments.

4.4.1.2 Model Development

The model uses standard normalization (Equation 3.3) and the NLPD loss function (Equation 4.1). The output layer is the same as the one described in Section 4.2 and the direct probability approach is therefore adopted. A hyperparameter tuner is used to determine the architecture. Since there are over 65 million training samples in the dataset, we only provide the tuner with a subset of this data. The tuner uses 1 million randomly selected samples from the training set and 200,000

randomly selected samples from the validation set. Each model trained by the tuner will run for 50 training iterations (or epochs) with a batch size of 4,096. Upon completion, the 10 best models are saved based on the validation loss values. All 10 models are evaluated (see Section 4.4.1.3), and the best performing one is used as a base architecture for full training.

4.4.1.3 Model Analysis

When comparing model to satellite densities, we use the mean absolute error (MAE) metric in percentage form. To assess the quality of the ML uncertainty estimates, we use CES (Equation 4.3). Although the model is explicitly predicting exospheric temperature, the statistics are computed after those temperatures are supplied to NRLMSIS 2.0. To obtain density, the following process is required. 1) MSIS-UQ predicts μ and σ for exospheric temperature at a particular location. 2) This distribution is sampled 1,000 times to extract samples that can be input to NRLMSIS 2.0. 3) These exospheric temperatures are interfaced with NRLMSIS 2.0 using the desired location and required model drivers. 4) If desired, μ and σ can be estimated from the density samples.

4.4.1.3.1 Comparison with NRLMSIS 2.0 and HASDM

To assess the validity of the model in terms of mean density prediction, its error with respect to the satellite estimates are compared to those of NRLMSIS 2.0 and HASDM. To get NRLMSIS 2.0 errors, the model is evaluated at all locations and times of the satellite measurements. For HASDM, the 3-dimension density grids from the SET HASDM density database are interpolated in log-scale to the satellite locations and times [57]. We then break up the errors into the three sets used for ML model development (training, validation, and test). We do this to simultaneously test the generalization of our model while ensuring differences in performance across the sets is also seen with the other models. In addition to the error assessment, we also compute the CES for MSIS-UQ across the three sets (in terms of density). For information on the conversion from ML predicted exospheric temperature to NRLMSIS 2.0 adjusted density, see Weimer et al. [69, 70].

4.4.1.3.2 Uncertainty Demonstration

The reliability of the MSIS-UQ uncertainty estimates is demonstrated early in Section 4.4.2, but the capabilities are further established in Section 4.4.2.1. The ML model directly predicts the uncertainty into the exospheric temperature which is then incorporated into NRLMSIS 2.0. The probabilistic T_∞ values result in probabilistic local temperatures and species densities. We consider a given epoch (May 13, 2007 at 21:42.50 UT) where CHAMP and GRACE are at very different locations; CHAMP is near the equator on the night-side while GRACE is at high latitude on the day-side. NRLMSIS 2.0 is provided probabilistic T_∞ values from the MSIS-UQ distribution at each location, and we consider the temperature, species densities, and mass density between 100 and 800 km altitude. The distributions are shown as a function of altitude, and the satellite estimates are provided for reference.

4.4.2 MSIS-UQ Results

Figure 4.14 shows the relative error distributions and mean absolute error for NRLMSIS 2.0, HASDM, and MSIS-UQ with respect to density estimates from CHAMP, GRACE, Swarm A, and Swarm B. The calibration curve for MSIS-UQ is also displayed alongside the calibration error score. This is separated by samples in the MSIS-UQ training, validation, and test sets. Similar figures are provided for each individual satellite in the Supplementary Materials.

Panel (a) shows the altitudes for each satellite used in this analysis showing over a 200 km span over 15 years of measurements. The left-most panels (b), (d), and (f) indicate that MSIS-UQ provides much more accurate density predictions than both NRLMSIS 2.0 alone and HASDM. All three models have a tendency to overpredict density although MSIS-UQ has the smallest bias. The MAE values highlight the $\sim$25% error reduction from NRLMSIS 2.0 and the $\sim$11% error reduction from HASDM. Across the three sets, MSIS-UQ is well-generalized with density prediction errors ranging < 1.5%. With respect to its uncertainty estimates (panels (c), (e), and (g)), MSIS-UQ has a CES < 5% across the three sets. It has a tendency to underestimate in the middle prediction intervals (between 20% and 80%) but is well-calibrated at prediction intervals > 90%.

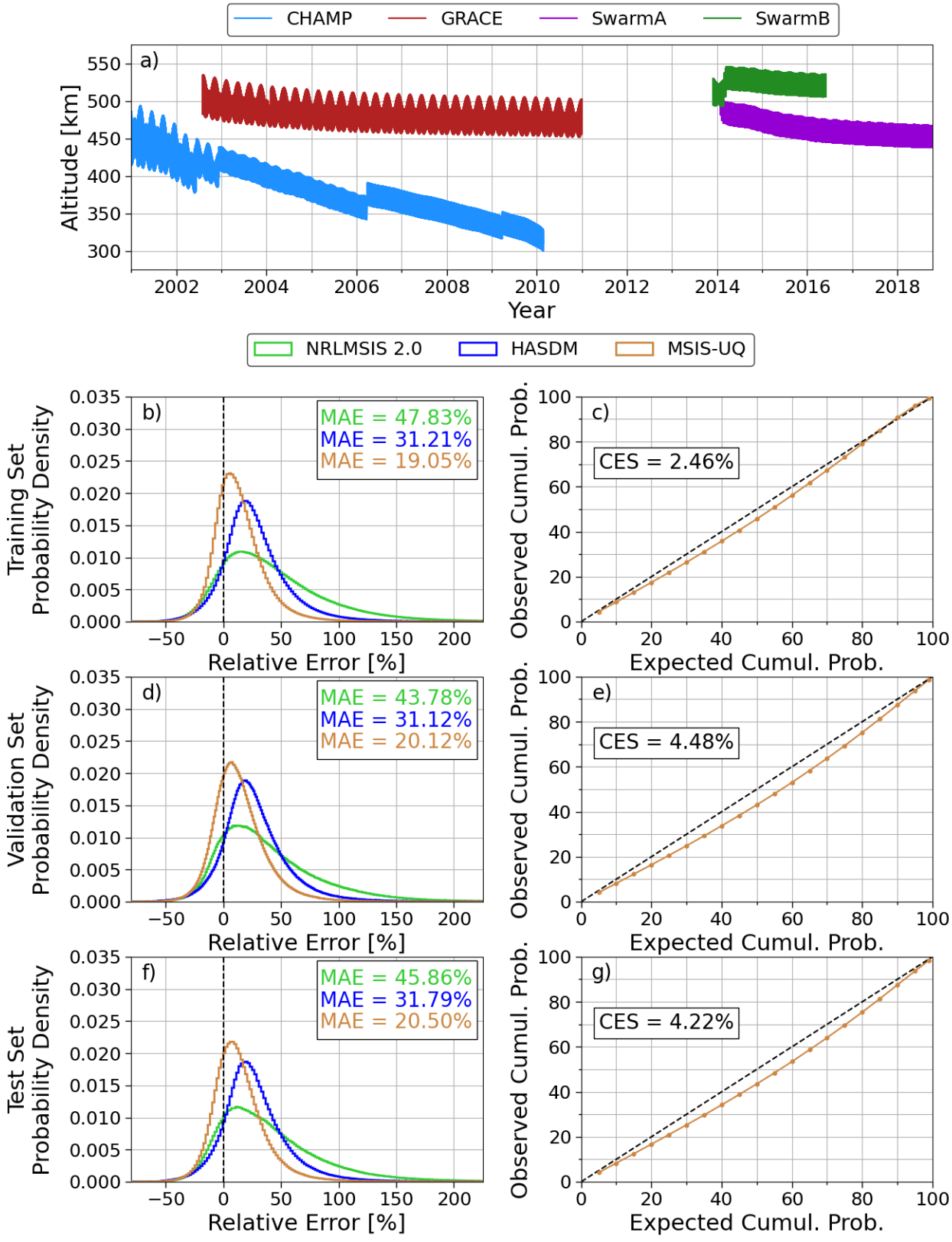

Figure 4.14: Altitudes of the satellites used for temperature and density estimates (a), relative error histograms (b,d,f), and MSIS-UQ calibration curves (c,e,g).

4.4.2.1 Uncertainties as a Function of Altitude

Figure 4.15 contains uncertainty profiles for MSIS-UQ at CHAMP and GRACE locations on May 13, 2007. There are panels for species density, temperature, mass density, relative uncertainty, and satellite position. Please reference the figure caption and Section 4.4.1.3.2 for details.

Panels (a) and (b) show the species density profiles at CHAMP and GRACE positions, respectively. The uncertainty bounds provide valuable information on the impact of exospheric temperature uncertainty on the uncertainty of local species. For example, one can investigate the Oxygen (O) to Helium (He) transition for various locations and conditions. Panel (a) shows that at CHAMP's position, this transition is occurring somewhere in the region of 507 to 552 km (1-σ) while at GRACE's position, the transition may occur between 688 and 738 km (1-σ). Other insights can be gained such as which species are most impacted by exospheric temperature at a given location/altitude. Note: only 1-σ bounds are shown here to prevent artifacts at low-values caused by the semi-logarithmic scale. The scale also causes the bounds to appear to be not-centered about the mean.

Panels (c), (d), and (e) provide information on the local temperature and mass density with uncertainty. In panel (c), MSIS-UQ severely shifts the exospheric temperature prediction and brings it closer to the estimates of CHAMP and GRACE; in both cases NRLMSIS 2.0 overpredicts temperature. The uncertainty in temperature is unobservable below 130 km and grows until it reaches the asymptotic temperature between 250 and 300 km. The uncertainty bounds and mean remain unchanged above these altitudes. Note that the CHAMP and GRACE temperature estimates are for T_∞ but we show them at their current altitude as the temperature has converged.

In panel (d), we see different trends in mass density. Again the uncertainty is minimal below approximately 200 km and begins to increase for a few hundred kilometers. The overprediction of temperature in NRLMSIS 2.0 results in higher than observed density by CHAMP and GRACE around 350 and 475 km, respectively. MSIS-UQ provides a more accurate density predictions at the satellite locations. Panel (e) shows the 1-σ uncertainty with respect to mean density. This shows different model behavior between the two locations. At CHAMP's location, the uncertainty

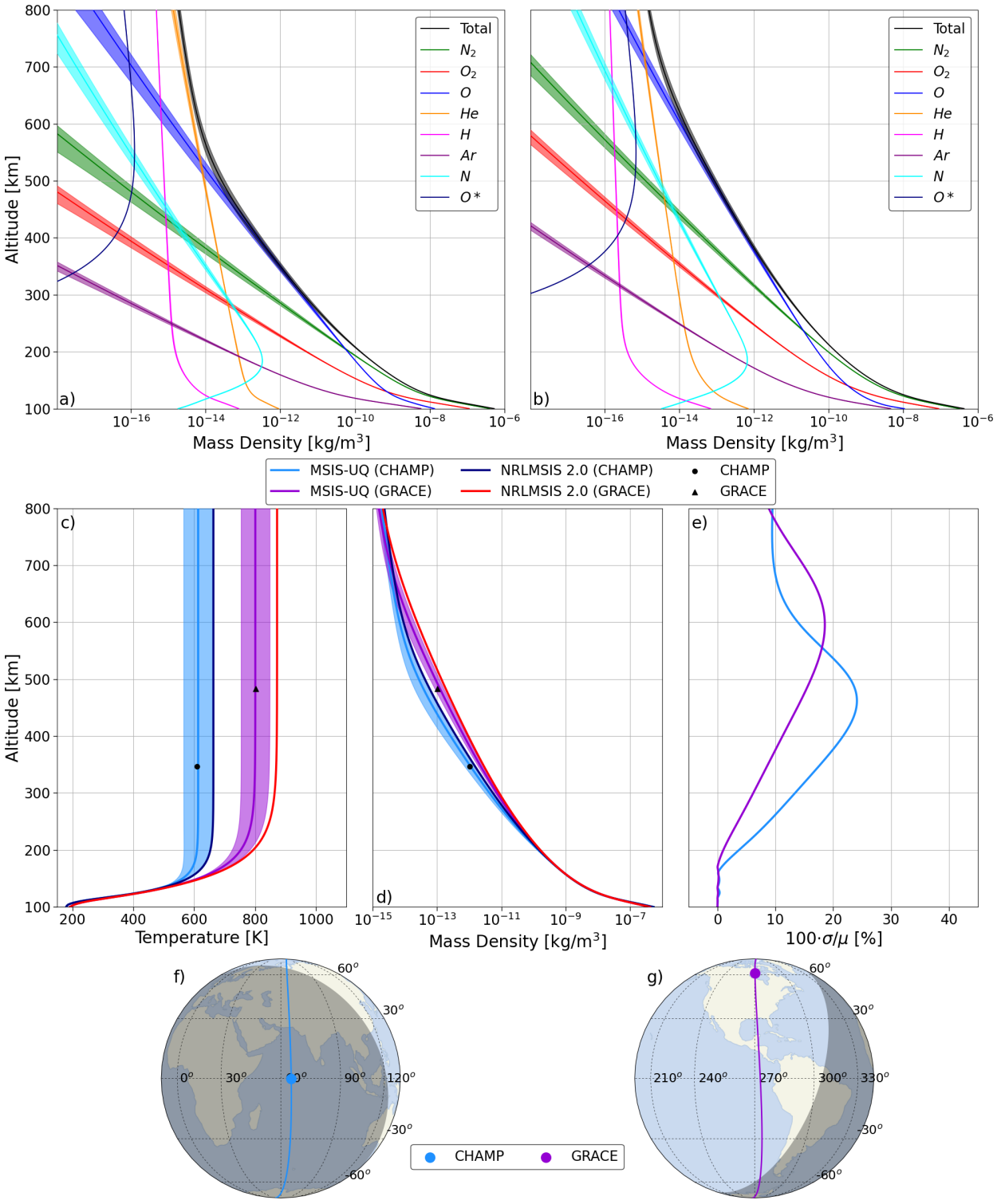

Figure 4.15: MSIS-UQ species density profiles for CHAMP (a) and GRACE (b) locations with 1-σ bounds, temperature profiles with 2-σ bounds (c), total mass density profiles with 2-σ bounds (d), 1-σ uncertainty normalized by the mean prediction (e), and the paths for CHAMP (f) and GRACE (g) with the current location denoted by the markers. This was conducted for May 13, 2007 at 21:42.50 UT.

increases to 24% around 460 km and decreased until around 700 km where it settles to 9%. At GRACE's location, the uncertainty continues to increase until it reaches 18% at 600 km where it begins to decrease.

4.5 TIE-GCM ROPE

Empirical and assimilative models (e.g. MSIS and HASDM) are useful in operations as their predictions are based on knowledge of the thermosphere combined with decades of observational data. However, there are no constraints to satisfy the many physical equations that describe the overall system in space and time. This is what drives the development of physics-based models. These satisfy the governing equations and provide a more realistic evolution of the thermosphere, particularly during storms. However, their computational expense and difficulty to incorporate uncertainty limit their usefulness in STM applications where uncertainties are vital, and there is a growing frequency of potential conjunction events. We attempt to leverage reduced order modeling, RNNs, and ensemble modeling to alleviate these drawbacks.

4.5.1 LSTM Methodology

Over the last several years, some researchers have developed dynamic reduced order models (ROMs) for empirical and physics-based thermosphere models alike. Mehta et al. [116] used PCA on TIE-GCM data and developed a dynamic ROM using dynamic mode decomposition (DMD) with control (or DMDc). This approach has been applied by Gondelach and Linares [117, 118] on the NRLMSISE-00, JB2008, and TIE-GCM models with the goal of data assimilation. DMDc is based on the assumption of the linear relationship between successive time steps and the processes that drive the system,

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k \tag{4.8}$$

where $\mathbf{x}$ denotes the state, $\mathbf{A}$ is the dynamic matrix, and $\mathbf{B}$ is the input matrix relating the system inputs/drivers and successive state of the system. Here, k refers to the time step. In a matrix

format, this is achieved by $\mathbf{x}_{k+1}$ being the PCA coefficients from $2:n$, $\mathbf{x}_k$ being the coefficients from $1:n-1$, and $\mathbf{u}_k$ being the chosen drivers from $1:n-1$. As this is used as a benchmark for ML dynamic model development, the reader is referred to Proctor et al. [119] for the theory behind DMDc and to Mehta et al. [116] for details on its application to this dataset.

4.5.1.1 Data Selection and Preparation

To develop TIE-GCM Reduced Order Probabilistic Emulator (TIE-GCM ROPE), TIE-GCM density outputs are required. Mehta et al. [116] developed an input set of F_{10} and *Kp* containing one year of simulated outputs – resulting in 8,760 hourly input values. For F_{10}, they used a sine wave with a period of one solar rotation (27 days) that had minimum and maximum values of 60 and 250 sfu, respectively. The *Kp* was randomly sampled from observed distributions. This input set and the resulting TIE-GCM density is referred to as "Sim1". This dataset essentially contains a solar cycle worth of density variations in a single year making model development easier. In addition to this simulated dataset, TIE-GCM was run for an entire solar cycle (1996–2008).

TIE-GCM is a dynamic model, meaning it models the evolution of the system. Both HASDM and MSIS are static models, meaning they only make predictions with the current epoch in consideration. This aspect of TIE-GCM makes it valuable in scientific studies and therefore a surrogate model should work the same way. To develop a dynamic ML model, we leverage Long-Short Term Memory neural networks (LSTMs, see Section 3.4). Like vanilla recurrent neural networks, a number of "lag steps" must be defined. For TIE-GCM ROPE, three lag steps are used. Since the TIE-GCM dataset has a cadence of one-hour, this corresponds to a three-hour window which is generally enough time for perturbations to the input to be seen throughout the thermosphere. The internal cell memory allows for the longer-term effects to be accounted for. The study in Chapter 5 will show that the data-driven CHAMP-ML model had the strongest relationship between density and either current geomagnetic indices or indices from the last three hours.

This dataset has the following spatial resolution: 24 local solar time values, 20 latitude values, and 16 altitudes evenly spaced between 100 and 450 km. As with HASDM, the spatial dimensionality of this dataset is too large for uncertainty quantification methods. Therefore, PCA is applied

($r = 10$) as described in Section 3.5. For the tuner, the Sim1 dataset was chosen for manageable run times, and three segments from the solar cycle were chosen for validation. The three 1,000 sample segments represent high geomagnetic activity (late 2003), solar minimum (mid-2008), and solar maximum (early 2002). The PCA coefficients and space weather drivers for this tuning and testing dataset are shown in Figure 4.16.

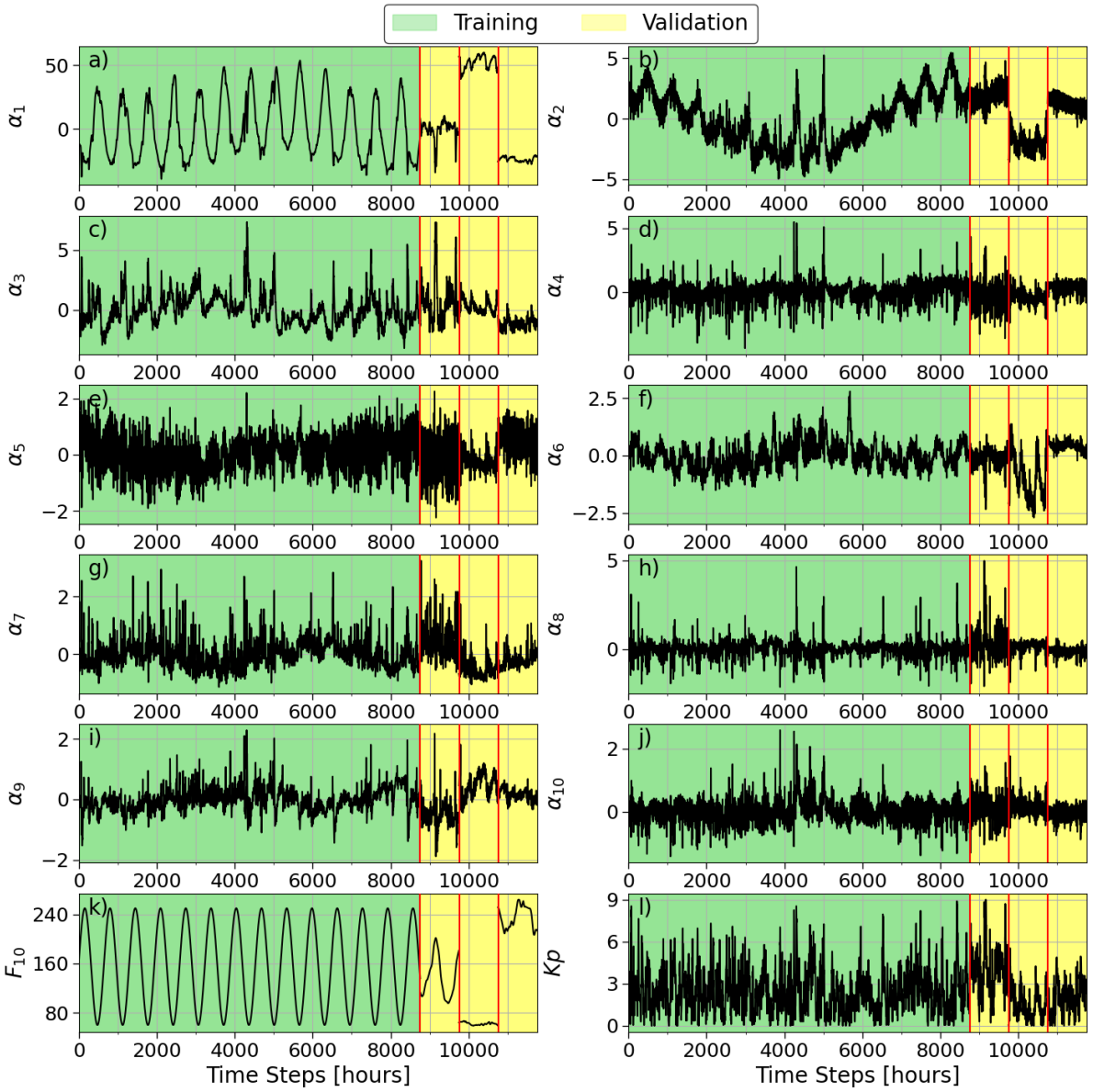

Figure 4.16: TIE-GCM PCA coefficients for Sim1 dataset and selected validation segments (a – j) with corresponding F_{10} (k) and Kp (l). The validation segments are shown as presented in the text (late 2003, mid-2008, early 2002).

4.5.1.2 Model Development

For other models developed in this work, direct probability distribution prediction resulted in both low error and robust uncertainty estimates. However, they were static, not recurrent, neural networks. Early efforts showed that the direct probability method was strongly underestimating uncertainty due to the one-step training and dynamic prediction process (Section 3.4.2). To attempt to combat this, we overhauled the default LSTM training process and force it to use a dynamic training process, mimicking its operational usage to get better predictions of σ. This required batch averaging and resulted in poor mean prediction capabilities. The final approach was to develop μ and σ models separately to try and leverage the benefits of both previous tests. The μ models could perform dynamic prediction with low error, but the σ models still struggled to provide meaningful uncertainty estimates. Another approach to uncertainty quantification, popular in terrestrial and space weather applications, is ensemble modeling [120, 121]. We therefore develop many individual LSTMs and combine the results in such a way to obtain accurate predictions and reliable uncertainty estimates.

Hyperparameter tuners were run on the Sim1 dataset to determine architectures for the aforementioned direct probability tests. While those did not yield an adequate final model, the last attempt (separate μ and σ modeling approach) provided architectures for potential ensemble models. The tuner options and search space used for the MSE models are provided in Table 4.14. These models were evaluated on the validation set shown earlier in Figure 4.16, and the top two architectures were used for future training.

Early modeling efforts showed that the Sim1 dataset, while valuable for testing and tuning, did not provide adequate model performance when used for training. Therefore, the training data for the final model comes from the TIE-GCM outputs spanning 2002–2008. This provides the model with the highest and lowest extremes of solar activity seen in the solar cycle. PCA is performed again on this time period, and all density data (1996–2008 and Sim1) is transformed using the basis functions from this training set. For validation, a block of 1,250 samples (approximately 52 days) from Sim1 is used as it provides a wide range of conditions for evaluation in a short period.

Table 4.14: Hyperparameter tuner parameters (left) and search space (right) for the mean square error LSTM.

Tuner Option	**Choice**	**Parameter**	**Values/Range**
Scheme	Bayesian Optimization	*Number of LSTM Layers*	1–3
Total Trials	50	*LSTM Neurons*	min = 32, max = 512, step = 4
Initial Points	25	*LSTM Activations*	tanh, sigmoid, softsign
Repeats per Trial	3	*Number of Dense Layers*	1–3
Minimization Parameter	val_loss	*Dense Neurons*	min = 64, max = 1024, step = 4
Epochs	2,500	*Dense Activations*	tanh, sigmoid, softsign, relu, elu, softplus
Early Stopping Criteria	val_loss	*Dense Dropout*	min = 0.01, max = 0.50, step = 0.01
Early Stopping Patience	75 epochs	*Optimizer*	RMSprop, Adam, Adadelta, Adagrad

One potential problem with this data is that there must be continuity for the LSTM internal state, but the model should not see the data in the same order every epoch – starting with solar maximum and ending with solar minimum. To avoid potential issues, the seven years are split into 490 segments with 125 time steps (approximately five days) of continuous data within them. The training process can be modified such that the model can be trained on each 125 sample continuous segment, and the internal state can be reset after each one. These 490 segments can be shuffled such that the LSTM sees different ordering of the data while being able to fine-tune its internal cell parameters without the threat of discontinuous data.

Another stark difference between Sim1 and 2002–2008 is the low relative frequency of geomagnetic storms in the historical period. Early testing also showed that the LSTMs trained on the historical data were more accurate overall compared to Sim1 models, but they had higher storm-time errors. A straightforward solution is to use sample weighting: applying an importance to each individual sample. We use a simple algorithm based on the frequency of samples at different *Kp*

levels such that the number of samples within a *Kp* bin multiplied by the number of samples in that bin is equal across all bins. This can help enforce importance to storm-time samples based on the relative frequency of these events.

At this point, the architectures for both models are finalized. To obtain the final μ model, it is trained on this new dataset with the described sample weighting scheme. At first, the model is trained using a typical one-step training method, but the loss is averaged over the 125 sample segment. This is performed for up to 2,500 epochs with early stopping based on mean absolute error in the density space for the validation set. After each validation segment, the true and predicted PCA coefficients are converted back to density through the inverse PCA transformation. Since the importance of the coefficients are not uniform, the MSE in the PCA space is not directly correlated to the best model. After the model is finished with batch training, it continues training without batch averaging using a smaller learning rate until the early stopping criteria is met once more. We obtain five models using this approach for the best two architectures resulting in ten models to make up the ensemble.

4.5.1.3 Weighted Averaging and Uncertainty Scaling

To derive the weighting arrays and uncertainty scaling factors, we do all computations separately within each architecture ($i = 1, 2$). The combination of the two architectures completes the ensemble, predicting different possibilities for a period of interest. Each individual model uses its own outputs as inputs (dynamic prediction), so the combination is done post-prediction.

The predictions of the PCA coefficients from each model will differ with varying levels of accuracy. Instead of simply averaging the predictions across the five models (for a given architecture), we opt to determine weighting factors based on the relative error of each model. To achieve this, each model is evaluated across the training set in five-day dynamic segments. The predictions of the PCA coefficients for each model and period are saved for later evaluation. The mean absolute error is computed for each model and each coefficient over the entire training set resulting in

a 5×10 array. Using this, the weights (w) are computed as,

$$w_{i,j,k} = \frac{\widetilde{w}_{i,j,k}}{\sum_{j=1}^{5} \widetilde{w}_{i,j,k}} \qquad where \qquad \widetilde{w}_{i,j,k} = \frac{1}{\text{MAE}_{i,j,k}} \tag{4.9}$$

where i, j, and k refer to each architecture, model, and PCA coefficient, respectively. $\widetilde{w}_{i,j,k}$ denotes the weights at an intermediate step before normalization. Once computed, the weighted mean and variance for each PCA coefficient from each architecture can be obtained,

$$\hat{\alpha}_{i,k,t} = \sum_{j=1}^{5} w_{i,j}\hat{\alpha}_{i,j,k,t} \qquad and \qquad \hat{\sigma}^2_{i,k,t} = \sum_{j=1}^{5} w_{i,j}\left(\hat{\alpha}_{i,k,t} - \hat{\alpha}_{i,j,k,t}\right)^2 \tag{4.10}$$

where $\hat{\alpha}_{i,k,t}$ and $\hat{\sigma}^2_{i,k,t}$ are the ensemble mean and variance for the i^{th} architecture and k^{th} PCA coefficient at time t, respectively. $\hat{\alpha}_{i,j,k,t}$ refers to the corresponding prediction from each of the five models. While this will provide a distribution under a Gaussian assumption, it does not guarantee robustness and reliability of the resulting uncertainty estimates. This can be improved with so-called σ scaling. Laves et al. [122] came up with a scaling factor (s) to scale model σ to better represent uncertainty. The scaling factor can be computed using the following equation, based on Equation 9 in [122].

$$S_{i,k} = \sqrt{\frac{1}{n_{tr}} \sum_{t=1}^{n_{tr}} \frac{\left(\alpha_{k,t} - \hat{\alpha}_{i,k,t}\right)^2}{\hat{\sigma}^2_{i,k,t}}} \tag{4.11}$$

In Equation 4.11, $S_{i,k}$ is the scaling factor for the i^{th} architecture and k^{th} PCA coefficient, n_{tr} is the number of time-steps in the training set, and $\alpha_{k,t}$ is the true/reference value for the k^{th} PCA coefficient at time t. The ensemble weights and scaling factors are saved for all later model use. The overall ensemble mean and variance can be computed as,

$$\hat{\alpha}_{k,t} = \frac{1}{2}\sum_{i=1}^{2} \hat{\alpha}_{i,k,t} \qquad and \qquad \hat{\sigma}^2_{k,t} = \frac{(n_1 - 1)\hat{\sigma}^2_{1,k,t} + (n_2 - 1)\hat{\sigma}^2_{2,k,t}}{n_1 + n_2 - 2} = \frac{\hat{\sigma}^2_{1,k,t} + \hat{\sigma}^2_{2,k,t}}{2} \tag{4.12}$$

where n_1 and n_2 are the number of models within each architecture. This is the pooled variance

formula which reduces to a simple average due to an equal number of models in each architecture. The final ensemble will be referred to as TIE-GCM reduced order probabilistic emulator (ROPE) due to its ability to provide Gaussian uncertainties and to function as a reduced order emulator for TIE-GCM (see Section 4.5.2.2).

4.5.1.4 DMDc Approaches

Given that considerable work has been done by other researchers for dynamic thermosphere modeling with DMDc, we use it as a baseline to compare with the LSTM. Gondelach and Linares [117] compared DMDc for NRLMSISE-00, JB2008, and TIE-GCM using linear and nonlinear inputs. For TIE-GCM, they used Kp^2 and $Kp{\cdot}F_{10}$ to try and overcome DMDc's limitation of linearity. To conduct a thorough test, we consider the LSTM test set of 1996–2001, and split the data into three segments using five-day dynamic prediction windows. (1) $Kp_{max} < 5$, (2) $5 \leq Kp_{max} < 7$, and (3) $Kp_{max} \geq 7$. Within these windows, we align them such that the maximum Kp is at the two-day mark.

We then create DMDc models with five different input sets based on TIE-GCM from 2002–2008 (the LSTM training set). All models use t_1–t_4 as temporal inputs (see Equation 2.1). The linear inputs are F_{10} and Kp, and the two nonlinear inputs are the ones used by Gondelach and Linares. Figure 4.17 shows the mean absolute error as a function of time for these models.

For the quiet case, panels (a,d), all DMDc models have similar errors. The average error across the five-day windows are within 0.6%. Although the periods are aligned with respect to the maximum Kp, the level of geomagnetic activity is low, and the error is therefore minimally affected. For the moderate case, panels (b,e), there is a sharper rise in error around the onset of a storm – nearly doubling in $\sim$12 hours. Still, there is little deviation between the models, and the mean error for all models is within 0.7% over five days.

The strong storms create stark differences in model performance. While the strictly linear approach has the highest errors at maximum Kp, it has the best recovery from the storm. Including the additional nonlinear inputs results in the worst post-storm performance which could be a result of using too many drivers, all changing drastically during these periods. Using only the two

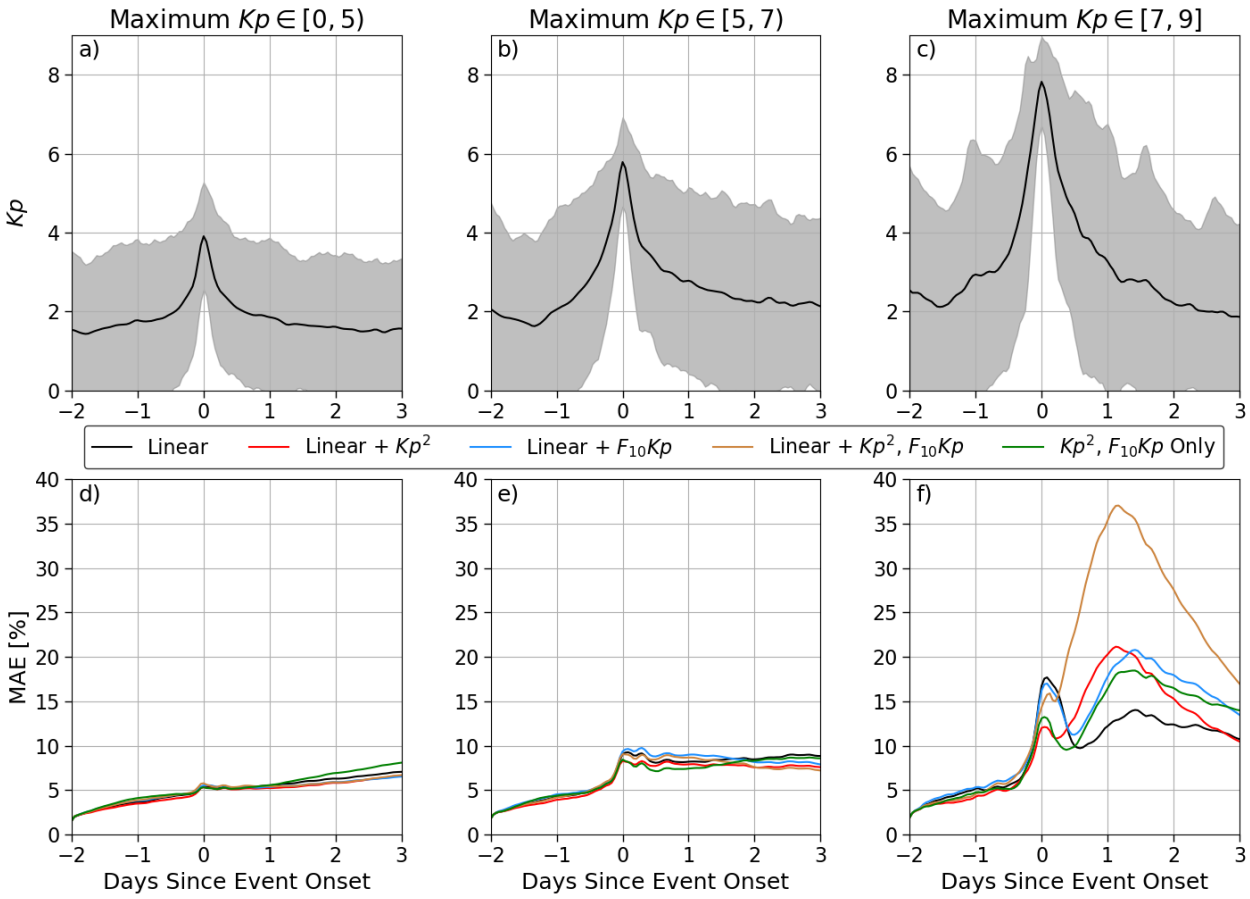

Figure 4.17: Average Kp for the three conditions with shaded 2σ bounds (a–c) with the corresponding errors (d–f). The legend denotes the models used in panels (d–f). Note: all models use the same temporal inputs.

nonlinear drivers (and time) does reduce the error at maximum Kp by about 5%, and the rise in error after the storm is not as pronounced. We proceed with the use of this DMDc approach due to storm-time improvement and use in other work. For Figures 4.19 and 4.20 in the proceeding section, we show both the linear and nonlinear-input DMDc models which will be referred to as DMDc and DMDc NL, respectively. All other tables and figures use the nonlinear-input DMDc model.

4.5.2 TIE-GCM ROPE Results

This dynamic reduced order modeling effort poses challenges when determining the best way to evaluate the models. With static modeling (e.g. HASDM-ML, CHAMP-ML, MSIS-UQ), error

and calibration can be analyzed relatively simply across the training, validation, and test sets. However, the DMDc and LSTM ensemble models will have different statistics depending on the evaluation window. We therefore must take careful consideration when evaluating and comparing the two methods.

4.5.2.1 Five-Day Operational Analysis

Once the final models were trained and the weighting and scaling schemes were determined (Section 4.5.1.3), the ensemble was evaluated on all available TIE-GCM data. The results for five-day dynamic prediction windows on the training, validation, and test sets is shown in Table 4.15 alongside the DMDc model. The calibration error score (Equation 4.3) is shown for TIE-GCM ROPE.

Table 4.15: Error and calibration statistics for DMDc and LSTM models averaged over 5-day dynamic prediction periods.

DMDc	**Set**	**Training**						
	Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
	MAE	6.31%	7.55%	6.81%	6.41%	6.02%	5.70%	5.97%
	Set	**Val.**	**Test**					
	Year	**Sim1**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**
	MAE	32.43%	5.00%	4.48%	5.43%	6.73%	7.01%	6.34%
LSTM	**Set**	**Training**						
	Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
	MAE	5.66%	6.44%	5.93%	6.87%	5.12%	7.23%	8.45%
	CES	16.16%	16.83%	15.08%	15.75%	15.56%	16.02%	18.49%
	Set	**Val.**	**Test**					
	Year	**Sim1**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**
	MAE	10.34%	5.57%	4.83%	5.60%	6.27%	6.38%	6.44%
	CES	13.53%	17.60%	15.13%	15.03%	15.16%	14.82%	15.90%

Across the training set (decline of solar cycle 23), both DMDc and the ensemble have similar errors, on the order of 4–9%. It is important to note that the errors are with respect to the TIE-GCM density, so it also includes an average of 2% truncation error from PCA. On the validation set, both modeling approaches have their highest error – 32% and 10% for DMDc and LSTM ensemble,

respectively. This is caused by the variability in the Sim1 dataset (Figure 4.16). There is a high concentration of storms, and solar activity changes the drastically on the time-scale of days. This causes some issues for the LSTM, an increase of about 4% error with respect to its average on the training set, but it causes the average DMDc error to jump 4–5 times its normal values for a given year. This will be explored further in Section 4.5.2.1.1.

On the test set, both models perform similarly to both each other and to their performance on the training set. This indicates good generalization on historical periods. The calibration error score for the ensemble is between 13% and 19% for any given year (including the validation set). While this is higher than in previous modeling efforts, dynamic prediction poses a challenge. Adding models and architectures to the ensemble could potentially reduce the CES. We test the robustness of the LSTM ensemble to geomagnetic activity by performing the study for DMDc (Figure 4.17) on TIE-GCM ROPE. The results are shown in Figure 4.18.

As seen in Figure 4.17, DMDc has low dynamic prediction error during geomagnetically quiet conditions. For this same period, the LSTM ensemble also has low errors, but the error climbs to $\sim 5\%$ within 24 hours while it takes DMDc around 72 hours to reach the same error. Beyond this point, the LSTM ensemble has lower errors. In panel (e), the error again climbs to 5% for the LSTM in the first day, but it is relatively unaffected by the onset of the storms. DMDc errors jump to around 8% and slightly increases post-storm while the ensemble only jumps to approximately 7% and drops to around 6% for the remainder of the dynamic prediction window.

For the strongest storms (panel (f)), TIE-GCM ROPE proves to be much more robust. The errors for it and DMDc converge around 24 hours after the dynamic prediction starts. At maximum *Kp*, the error for TIE-GCM ROPE peaks at just above 10% which is below even the lowest error at max *Kp* for any nonlinear DMDc approach tested in Figure 4.17. The LSTM error also continues to decrease post-storm back to the 7%–9% range. The 2σ error bounds to TIE-GCM ROPE area also considerably lower than for DMDc for the storm and post-storm periods. The calibration error score for TIE-GCM ROPE is also fairly consistent across the three conditions.

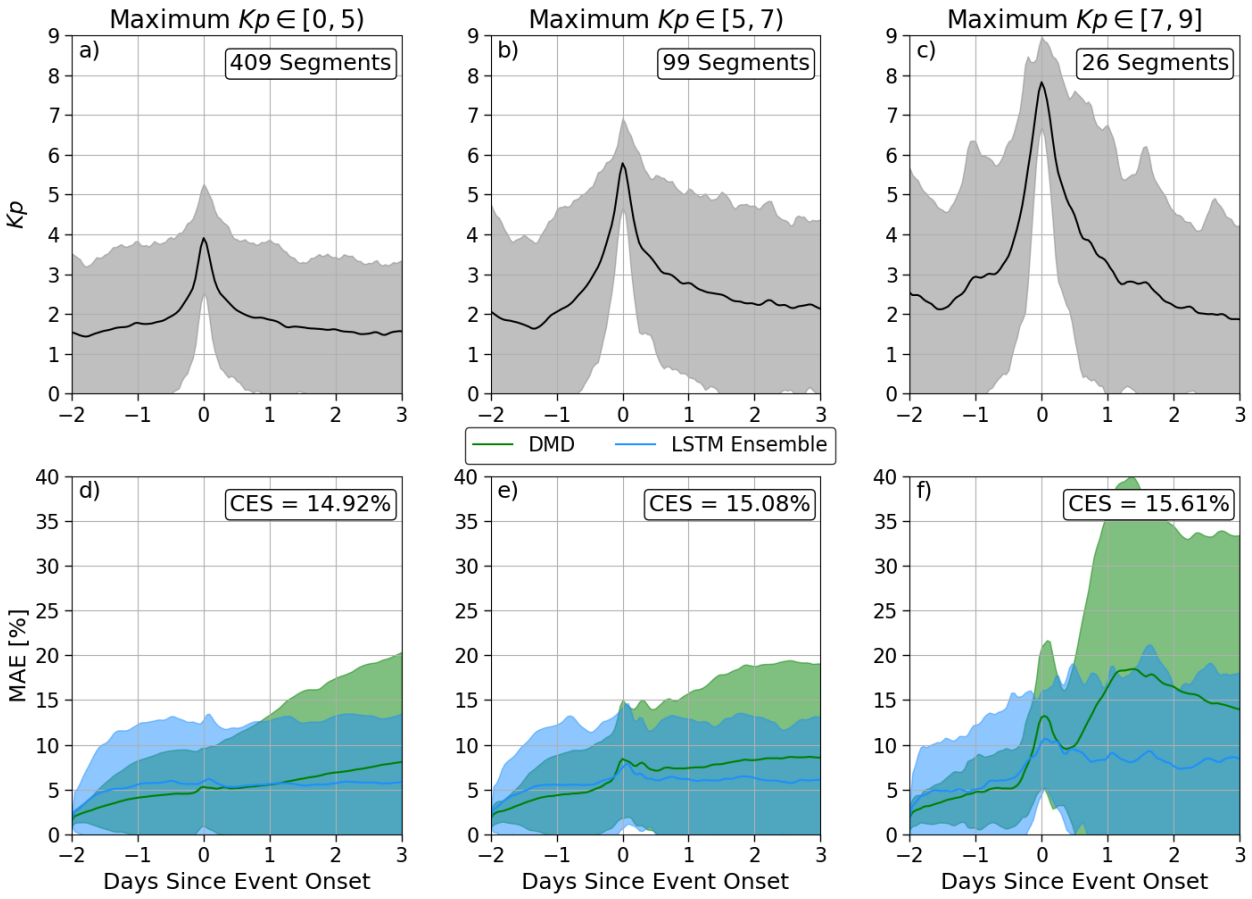

Figure 4.18: Average *Kp* for the three conditions (a–c) with the corresponding errors for DMDc and LSTM ensemble(d–f). The shading represents 2σ bounds for *Kp* (a–c) and errors (d–f).

4.5.2.1.1 DMDc Sensitivity

Figure 4.18 highlighted the robustness of TIE-GCM ROPE to geomagnetic activity while also showing the low error for DMDc during quiet periods. However, we noted in Table 4.15 that the five-day forecast errors for DMDc were significantly higher on the Sim1 validation data. While Sim1 contains an above-average number of geomagnetic storms for a year-long period, another distinguishing feature of the dataset is how drastically F_{10} can vary in a few days. We explore this further by considering a period from Sim1 that is outside the validation set. The DMDc and TIE-GCM ROPE predictions in the reduced state are displayed in Figure 4.19. Note: although we extract σ from the LSTMs to get distribution statistics for TIE-GCM ROPE, the individual model

predictions are shown alongside the ensemble mean for observational purposes. We also show the linear-input DMDc model for comparison.

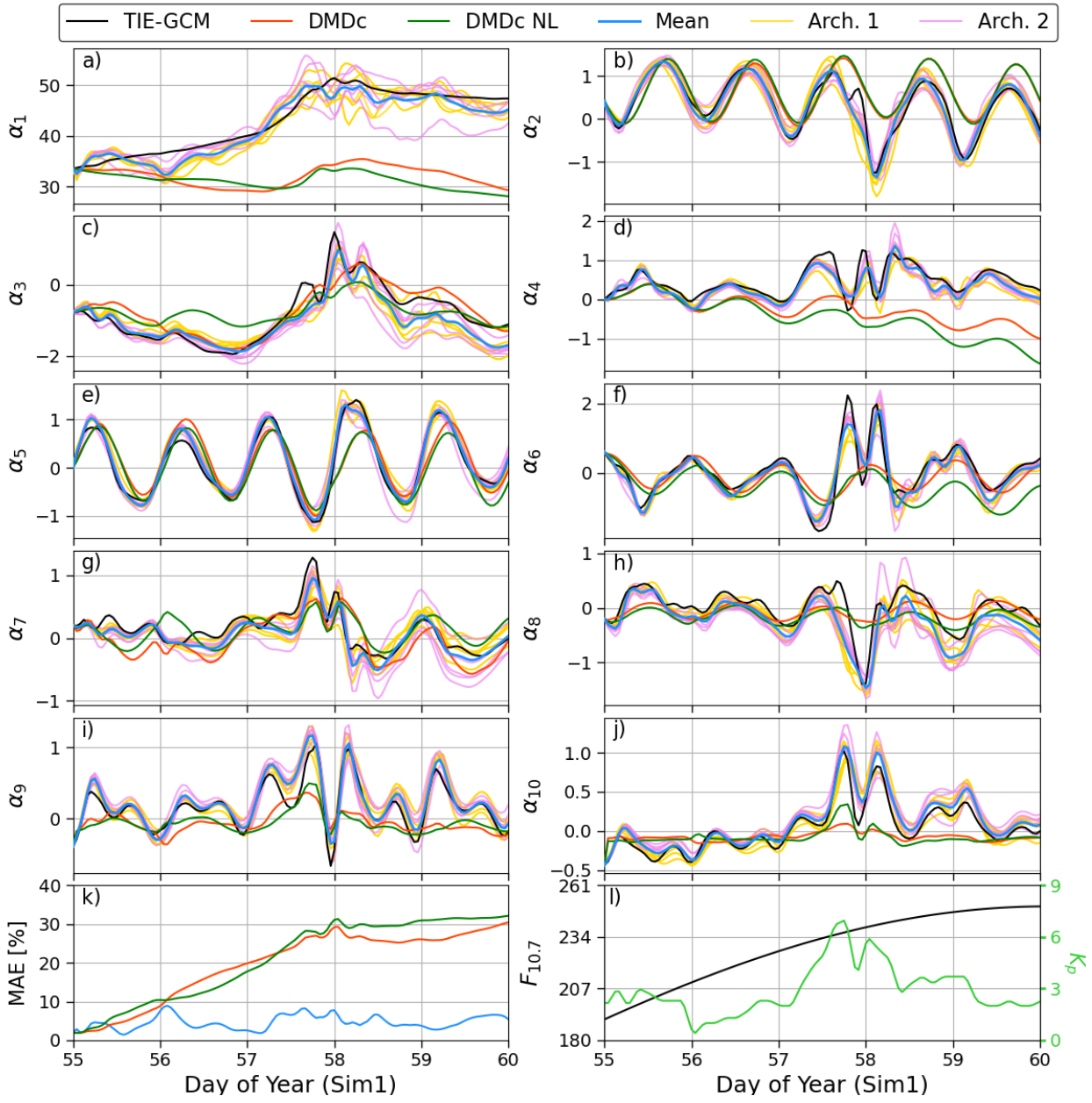

Figure 4.19: PCA coefficients from TIE-GCM along with dynamic prediction from the linear-input DMDc model, the nonlinear-input DMDc model (DMDc NL) and TIE-GCM ROPE (a–j), global density mean absolute errors for the two models (k), and corresponding space weather drivers (l).

Before discussing the DMDc and LSTM predictions in Figure 4.19, we must look at F_{10} and *Kp* for this period (panel (l)). F_{10} increases from 190–250 sfu in this five-day window. This is a substantial increase for this short of a time – which is very unlikely – but it provides us with a test for how well these models can function as a reduced order model for TIE-GCM. *Kp* is fairly quiet for the beginning of this period but rises to a moderate storm just before the three-day mark.

The most glaring result from Figure 4.19 is the DMDc predictions in panel (a). While the individual LSTMs properly track the quickly rising α_1 from TIE-GCM, the two DMDc models do not. In fact, it decreases until the storm onset at day 57. Considering that α_1 is the most important PCA coefficient (capturing the most variance in the dataset), this explains the larger DMDc errors in panel (k). While DMDc has generally low errors (Table 4.15 and Figure 4.18 panel (d)) it does not seem to follow large variations in the thermosphere, which is a major attribute of the Sim1 dataset. Conversely, TIE-GCM ROPE does well following the variations in all of the PCA coefficients, and the onset of the storm leads to some divergence in the individual LSTM predictions. This is expected as the variance should rise with storms, especially considering that it does affect the model performance (Figure 4.18). The ability for TIE-GCM ROPE to follow even the higher-order PCA coefficients shows that it can represent more of the dynamics in TIE-GCM relative to DMDc. It is also important to note how the distributions for each architecture evolve in different ways. For α_1 specifically, the second architecture results in more diverse model predictions following the storm, highlighting the importance of the hierarchical development of the ensemble.

4.5.2.2 Ensemble Emulation

To test emulation capabilities, the DMDc model and TIE-GCM ROPE were evaluated on the same periods, but with a dynamic prediction window of approximately one year. This is displayed in Table 4.16. Note that for 1996, there is only 280 days available for prediction, and the validation period on Sim1 is only 54 days.

For the long-term dynamic prediction, DMDc becomes unreliable with errors ranging from 38%–82%. For most years, the TIE-GCM ROPE error is similar to the corresponding values in Ta-

Table 4.16: Error and calibration statistics for DMDc and LSTM models averaged over full-length dynamic prediction periods. This is 280 days for 1996, 362 days for all other years, and 52 days for the validation set.

DMDc	**Set**	**Training**						
	Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
	MAE	61.94%	45.79%	49.11%	55.90%	59.36%	75.76%	81.30%
	Set	**Val.**	**Test**					
	Year	**Sim1**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**
	MAE	44.46%	82.02%	44.10%	40.21%	42.90%	52.24%	38.46%
LSTM	**Set**	**Training**						
	Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
	MAE	6.02%	6.75%	6.72%	8.14%	7.81%	12.94%	23.12%
	CES	15.78%	24.09%	26.48%	27.63%	28.52%	26.82%	22.28%
	Set	**Val.**	**Test**					
	Year	**Sim1**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**
	MAE	11.27%	18.70%	9.67%	6.21%	6.76%	6.73%	6.76%
	CES	5.59%	30.06%	29.29%	25.39%	17.30%	12.21%	15.11%

ble 4.15. It appears that the ensemble can emulate TIE-GCM for long periods with low errors with the exception of solar minimum (1996, 2007, 2008). During these periods, the errors are between 13% and 23%. The calibration error score for the ensemble is generally higher for these long-term dynamic prediction windows. To visualize the long-term dynamic prediction performance, we look at a 362-day prediction on the Sim1 dataset. The validation set is contained within this period but only accounts for $\sim$15% of the samples. The mean density at 400 km for TIE-GCM is shown with DMDc and TIE-GCM ROPE predictions in Figure 4.20.

While it is difficult to see, both DMDc and TIE-GCM ROPE have low errors for the first few days, but as seen with Figure 4.19, the DMDc error quickly compounds. The DMDc predictions do not follow the trends of TIE-GCM, but the cyclic nature of F_{10} in Sim1 allows for the DMDc error to briefly drop at times (panel (b)). However, the large errors, peaking above 1000%, require panel (b) to be shown in a logarithmic scale. TIE-GCM ROPE is able to track the long and short-term variations without any state updates. Its errors peaks at 30% but generally remains around 10% across the year-long prediction window. The individual LSTM predictions are more prominent

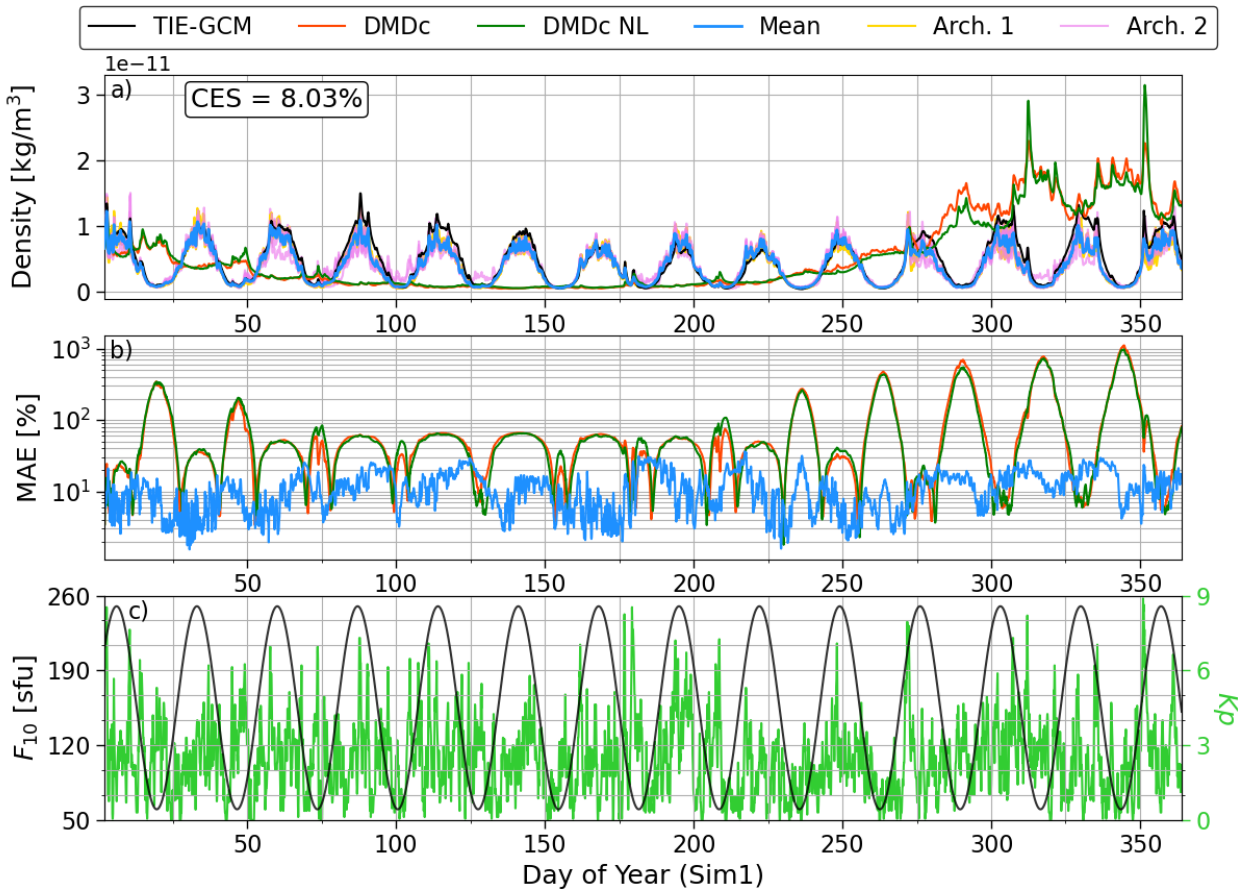

Figure 4.20: Mean density at 400 km (a) with global-averaged errors for linear-input DMDc, nonlinear-input DMDc (DMDc NL), and LSTM ensemble (b), and the corresponding space weather drivers (c) for a 362-day period across the Sim1 dataset.

during high-density levels in solar maximum conditions (see panel (c)), and the uncertainty is therefore higher in those conditions.

4.6 Summary

This chapter focused on the development of four unique probabilistic thermosphere models based in ML. Sections 4.1 and 4.2 discuss the evolution of HASDM-ML development. HASDM-ML is based on the state-of-the-art HASDM system which is not publicly available. These sections introduce the concept of MC dropout as a means for UQ along with the introduction of different loss functions that can help the model achieve robust and reliable uncertainty estimates (NLPD and CRPS). This transitioned into a more computationally efficient UQ technique: directly prediction

the probability distribution. The best HASDM-ML model resulted from the direct probability method and the NLPD loss function. We show that HASDM-ML is much more similar to HASDM than JB2008, the base model of HASDM, and HASDM-ML proves to be robust during extreme events.

CHAMP-ML was developed as a byproduct of the uncertainty quantification studies for HASDM as it is instead based on in-situ measurement. CHAMP-ML is able to learn the spatial relationship of the thermosphere in addition to its response to space weather events. We were able to study the uncertainties of CHAMP-ML to show that the model uncertainty is structured based on the dataset it was trained on. For example, CHAMP-ML had more uncertainty at 450 km during solar minimum than at 300 km, because at solar minimum conditions, CHAMP was at low altitudes.

MSIS-UQ, like CHAMP-ML, is based on it-situ satellite data; however, it predicts exospheric temperature as a standalone model as opposed to mass density. Nevertheless, MSIS-UQ functions as a thermospheric density model, because of its use as a coupled model with NRLMSIS 2.0. We showed that it could significantly reduce the bias between NRLMSIS 2.0 and satellite density data while simultaneously making MSIS probabilistic with robust uncertainty estimation capabilities. The combination of MSIS-UQ and NRLMSIS 2.0 provide a unique opportunity to study uncertainties in the composition of the thermosphere as well.

The final model developed was TIE-GCM ROPE, based on a dynamic physics-based thermosphere model. The use of LSTMs allowed for the ML model to model the dynamic nature of the thermosphere, but it created challenges with respect to UQ. The ensemble approach was chosen to bypass the poor performance with NLPD, and two scaling/weighting schemes provided increased accuracy and improved UQ performance. The studies for TIE-GCM ROPE showed that during storms, the nonlinear ML approach outperformed the linear DMDc method, and it is able to capture both the short and long-term dynamics of the original TIE-GCM model. These models will be used to show that they have a place in scientific studies for space weather (Chapter 5). They will also crucially show the importance of uncertainty of using model uncertainty in STM applications (Chapter 7).

Chapter 5. Science through Machine Learning

In the space weather community, ML has been used to develop models for problems such as solar flare prediction [123], ionospheric scintillation detection [124], and geomagnetic index forecasts [125]. However, its use is often limited to problem solving, not for investigative purposes. Further, there are rarely any studies on what the model has learned outside of determining its performance metrics. Convolutional neural networks – specifically related to image processing – are inherently easier to understand, as the filters (or weights) can be displayed as images and therefore interpreted [126]. This is not a luxury associated with ML regression models where inputs do not have visual qualities, so the weights are difficult to interpret. This motivates the work of this chapter as it pertains to space weather and the thermosphere.

5.1 Thermospheric Overcooling Phenomenon

Many thermosphere models struggle to quantify the amount of heating during a geomagnetic storm. Meanwhile the timing and severity of post-storm cooling remains a challenge for the modeling community. Zesta and Oliveira [127] found that when storms become stronger, the thermosphere both heats and cools at a faster rate. Significant research has been done into a potential cause of the cooling effects, overproduction of nitric oxide (NO) and its infrared emissions. Kockarts [14] investigated the cooling impact of the thermosphere due to downward heat conduction, atomic oxygen (O), and NO during a geomagnetic storm in 1974. They found that the reduction in thermopause temperature from the introduction of NO cooling was 440 K, while the addition of O cooling only reduced the temperature by another 35 K. This topic has gained much more attention in recent years due to the NO emission data from the Sounding of the Atmosphere using Broadband Emission Radiometry (SABER) instrument [128] and high fidelity density estimates from satellite such as CHAMP and GRACE.

Mlynczak et al. [15] used SABER data during the storm periods of April 2002 and found that NO emissions were notably enhanced during this period. Lei et al. [129] considered the

prominent 2003 Halloween storms to provide a comparison of SABER data to density estimates from both CHAMP and GRACE. They noted a 23–26% maximum density depletion during the recovery phase for the satellites relative to quiet pre-storm values, and the NO cooling rates during this period remained at a high level. Knipp et al. [130] examined 192 geomagnetic events to compare NO and neutral density data from GRACE. Their data-based study suggests shock-led interplanetary coronal mass ejections result in an overproduction of NO which provides a cooling effect that compensates for the strong thermospheric expansion that occurs during these storms. The driving force behind the cooling effect is still an active area of research and other mechanisms (e.g. ionosphere-driven atomic oxygen reductions) have been proposed to explain the phenomenon [131, 132]. We do not attempt to confirm any driving mechanisms in this work.

Using ML, we can investigate the *presence* of post-storm cooling in various datasets and which model drivers may be required to capture it. We first explain the data and models used for model development and comparison. Then, we describe the ML models and how we use them to examine this phenomenon. We show model predictions during a prominent geomagnetic storm to motivate the importance of this work and provide a quantitative analysis on the effect of geomagnetic time history on the predicted density.

5.2 Data, Models, and Methods

5.2.1 Data and Models

As a benchmark, we use NRLMSIS 2.0. As described in Section 2.3.1.1, it uses time series *ap* to account for geomagnetic activity over the previous 57 hours. This has the potential to inform the model of a recent strong storm. Machine-learned density models are developed based on four separate datasets. The first three ML models are HASDM-ML, CHAMP-ML-v2, and MSIS-UQ (all from Chapter 4). Note that all CHAMP-ML use for the remainder of this work is with CHAMP-ML-v2. The last model is developed on outputs of JB2008 from the start of 2000 to end the of 2019 [6]. JB2008 was evaluated every three hours and at a fixed grid of 12,312 locations including altitude. The space and time resolution is consistent with the SET HASDM density database. The

model drivers for JB2008 are described in Section 2.3.1.2. The ML drivers and model development methodology is consistent with that of the direct probability HASDM-ML model. This will be referred to as JB2008-ML in this chapter. It is important to note that TIE-GCM ROPE is not used here as it does not take time-series inputs required to conduct this analysis (see Section 5.2.3).

5.2.2 Storm Example

To motivate the work, we evaluate NRLMSIS 2.0 and the ML models from the various datasets during the 2003 Halloween storms. The time series *ap* flag was enabled when running NRLMSIS 2.0 in this work. The five models were provided the true drivers for the six day period from October 28 – November 3, 2003 and were compared to the Mehta et al. [48] CHAMP estimates. For NRLMSIS 2.0, CHAMP-ML, and MSIS-UQ, the predictions were made with the same time cadence and at the specific locations of the satellite, negating the need for further processing. For JB2008-ML and HASDM-ML, the models were evaluated at the 3-hour intervals used by the original models. The global density grids were then interpolated in space and time in log-scale to the locations of CHAMP. The final step is to take a running average of the along-orbit densities over the 92.5 minute orbital period to obtain orbit-averaged densities. This allows us to visualize the general density along the orbit during the storm period (Figures 5.1 and 5.3).

5.2.3 Time Lag Study

As discussed early in the chapter, cooling mechanisms often cause post-storm densities to be anomalously low. For this storm in particular, Lei et al. [129] noted nearly a 25% decrease in post-storm densities relative to pre-storm levels. In an effort to quantify this mechanism within the original models/datasets, the time histories for *ap* or *SYM-H* were independently varied within the models at four locations listed in Table 5.1. Table 5.1 also contains the geomagnetic indices held constant in each model while either *ap* or *SYM-H* were changed. All cases were at a constant solar activity with drivers set to 120. The time inputs were at 0 hours UT and represent the fall equinox (doy = 264), so there are no effects from Earth's tilt.

Table 5.1: Information for the time lag study. For clarification, LAT is latitude and S refers to both S_N and S_S.

Locations			
Night Equator	**Day Equator**	**Night Pole**	**Day Pole**
LST = 2 hrs, LAT = 0°	LST = 14 hrs, LAT = 0°	LST = 2 hrs, LAT = 80°	LST = 14 hrs, LAT = 80°
Constant Inputs			
NRLMSIS 2.0	**JB2008-ML**	**HASDM-ML**	**CHAMP-ML**
ap = 56	*ap* = 56, *Dst* = -50	*ap* = 56, *Dst* = -50	*SYM-H* = -50, *S* = 200

5.2.3.1 Additional Considerations for MSIS-UQ

Performing this study with MSIS-UQ requires additional consideration considering the coupling between it and NRLMSIS 2.0. To start, all non-geomagnetic model drivers are kept to constant values. We set the solar indices to 120 solar flux units, and the time for the study is 00:00 UT during the fall equinox. Each of the time-history geomagnetic drivers will be increased individually while all others are kept at a constant value: *ap* = 56, *SYM-H* = -50 *nT*. Since the ML model uses Poynting Flux totals and ΔT at epoch, they are kept constant at 200 GW and 120 K, respectively. MSIS-UQ uses *SYM-H* while NRLMSIS 2.0 uses *ap* for time-series geomagnetic drivers. To account for this distinction, we first fit a line between all *SYM-H* and *ap* values within our dataset. Using this, we find the *SYM-H* value associated with the *ap* value that must be used to get density from NRLMSIS 2.0. Therefore, density ratios for MSIS-UQ use this simultaneous *SYM-H* and *ap* variation as opposed to only using *ap* variations with NRLMSIS 2.0 alone. The results for MSIS-UQ will be presented separately to highlight the change from standalone MSIS.

5.3 Results and Discussion

We first show the error statistics for the four ML models in Table 5.2. These were computed with respect to the original datasets (e.g. JB2008-ML is the error with respect to JB2008 density). Training data is used to fit the model, validation data is used to determine the best model, and the independent test set used to determine performance.

Table 5.2 shows that JB2008-ML undoubtedly has the lowest errors, but it is worth noting that

Table 5.2: Mean absolute error on the training, validation, and test sets.

Model	Training	Validation	Test
JB2008-ML	5.28%	6.03%	6.63%
HASDM-ML	9.13%	10.46%	10.39%
CHAMP-ML	10.97%	11.60%	11.57%
MSIS-UQ	19.05%	20.12%	20.05%

it is also the most generalized dataset of the four. The SET HASDM density database contains evidence of more complicated processes and its PCA coefficients are more difficult to model as a result [113]. The CHAMP-ML errors are 1–2% higher than those of HASDM-ML, and MSIS-UQ has the highest errors with respect to its original dataset at around 20%. CHAMP-ML and MSIS-UQ are the only ML models in this work with location as a predictor. Furthermore, MSIS-UQ attempts to predict the exopsheric temperature to force NRLMSIS 2.0 to match the satellite density estimates which is a more challenging task. The HASDM-ML and CHAMP-ML error statistics differ from the previous chapter due to an updated architecture for HASDM-ML and the use of CHAMP-ML-v2. To both visualize the model performance in an operational setting and motivate the remainder of the work, we show the orbit-averaged densities for NRLMSIS 2.0, HASDM-ML, CHAMP-ML, and JB2008-ML compared to the Mehta et al. [48] CHAMP densities for the 2003 Halloween storms in Figure 5.1.

Figure 5.1 (a) shows that all models match the timing observed by CHAMP during both storms (10/29–10/30 and 10/30–10/31). NRLMSIS 2.0 has a tendency to overpredict density throughout this 6-day period, most notably between the two storms and in the recovery phase (11/02–11/03). This will be explored further with Figure 5.2. JB2008-ML exhibits similar behavior although it is closer to matching the contraction of the atmosphere between the storms. While both of these models use time-histories of *ap*, they do not portray any evidence of post-storm cooling. In contrast, HASDM-ML and CHAMP-ML both show significant decreases in density both between and after the storms. Oliveira and Zesta [133] performed a superposed epoch analysis showing density ratios for CHAMP/GRACE and JB2008 for extreme events which showed that JB2008 is

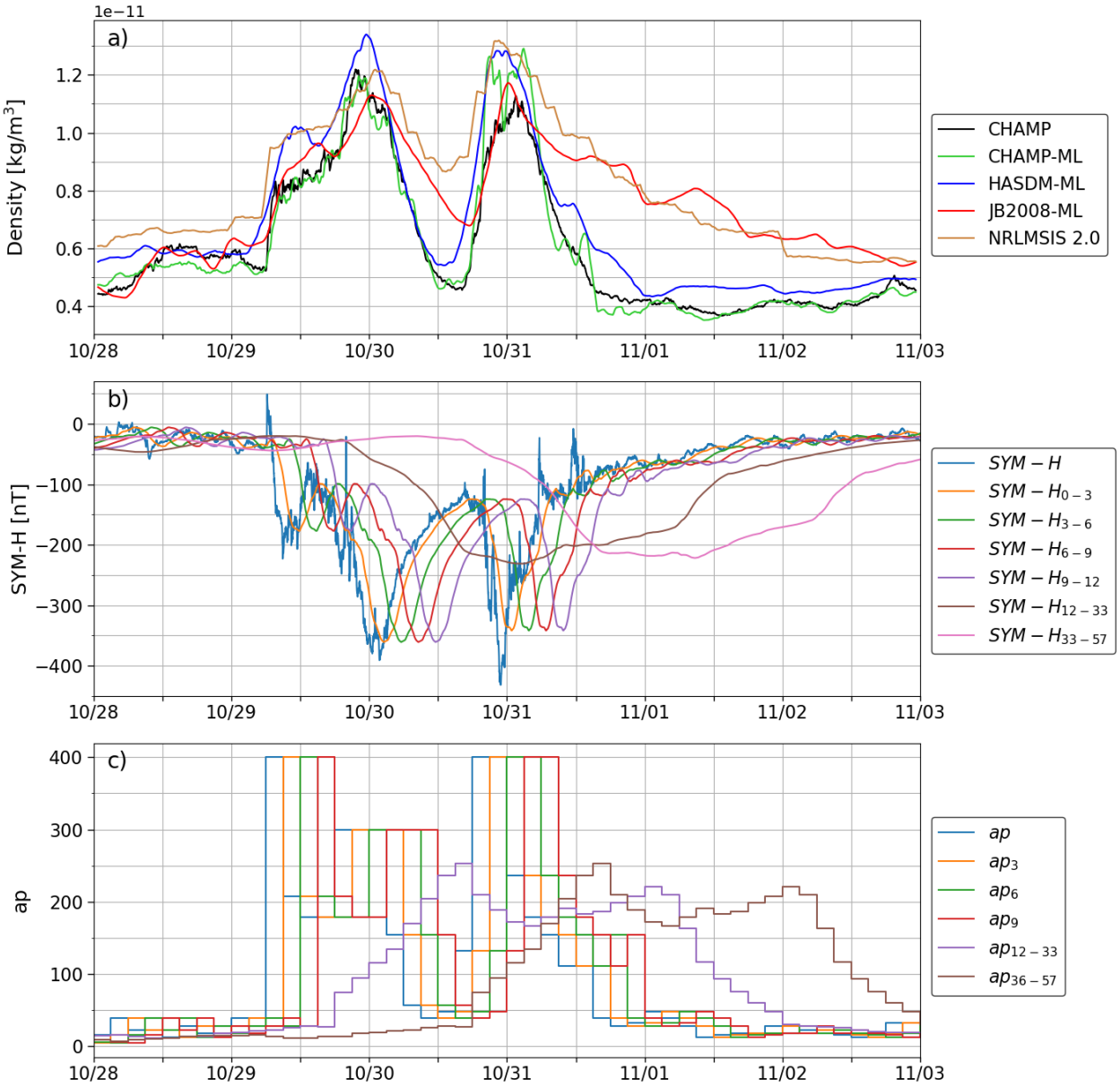

Figure 5.1: Orbit-average density for NRLMSIS 2.0, JB2008-ML, HASDM-ML, CHAMP-ML, and CHAMP (a) and the associated *SYM-H* (b) and *ap* (c) time-series inputs.

not modeling the satellite observed cooling taking place shortly after these events. In Oliveira et al. [134], this was expanded to include HASDM. While there were differences between the satellites and HASDM, the sudden cooling was consistent with observations.

Figure 5.1 (b) and (c) show *SYM-H* and *ap* time histories, respectively. The increased temporal

resolution for *SYM-H* is very evident, and the first four averages can inform CHAMP-ML of the recent magnetic disturbances. The last two time history inputs ($SYM\text{-}H_{12\text{-}33}$ and $SYM\text{-}H_{33\text{-}57}$) represent longer-term information with less variation. Immediately following the second storm (around 0600 UTC on 10/31), the last two time history inputs have large magnitudes while the more recent inputs no longer signify a storm. At the same time, the density predicted by CHAMP-ML and observed by the satellite drop abruptly. This behavior is consistent with the observation from Zesta and Oliveira [127] where CHAMP density experienced sudden decreases following extreme events. The *ap* time history is valuable to the other three models, following similar trends to panel (b) but are much more coarse. The results from the time-lag study (described in Section 5.2.3) are displayed in Figure 5.2.

Figure 5.2 is informative into what historical information is most important to represent the original data source – JB2008 output, SET HASDM density database, and CHAMP density estimates. NRLMSIS 2.0 is used here as a baseline due to its wide use in the field and use of historical geomagnetic information. There is a fairly linear relationship between the different *ap* values and density for NRLMSIS 2.0. In most cases, it considers the most recent *ap* to be most important and the least recent *ap* to be the least important. There is almost a perfect decay of slopes as it considers information from further in the past. At no point does the density ratio at the four locations drop below 1.00, which represents lower density than the baseline, or $ap_x = 0$ where x represents a given time-lag or lack thereof. This behavior explains why NRLMSIS 2.0 does not capture the behavior observed by CHAMP in Figure 5.1. The only drivers for the model that could capture post-storm behavior are the time history *ap* inputs. If NRLMSIS 2.0 could model post-storm overcooling, panels (a)–(d) in Figure 5.2 would have to have one or more curves that fall below 1.0 indicating its internal parameters account for this phenomenon.

For JB2008-ML there is virtually no evidence of post-storm cooling being present in the dataset. With the exception of the ap_9 curves, there is a fairly linear dependence between *ap* and density. Interestingly, JB2008-ML indicates that the strongest relationship between *ap* and density has a 9-hour delay. The ap_9 curves are nonlinear for $ap < 100$ and quite linear for $ap > 100$.

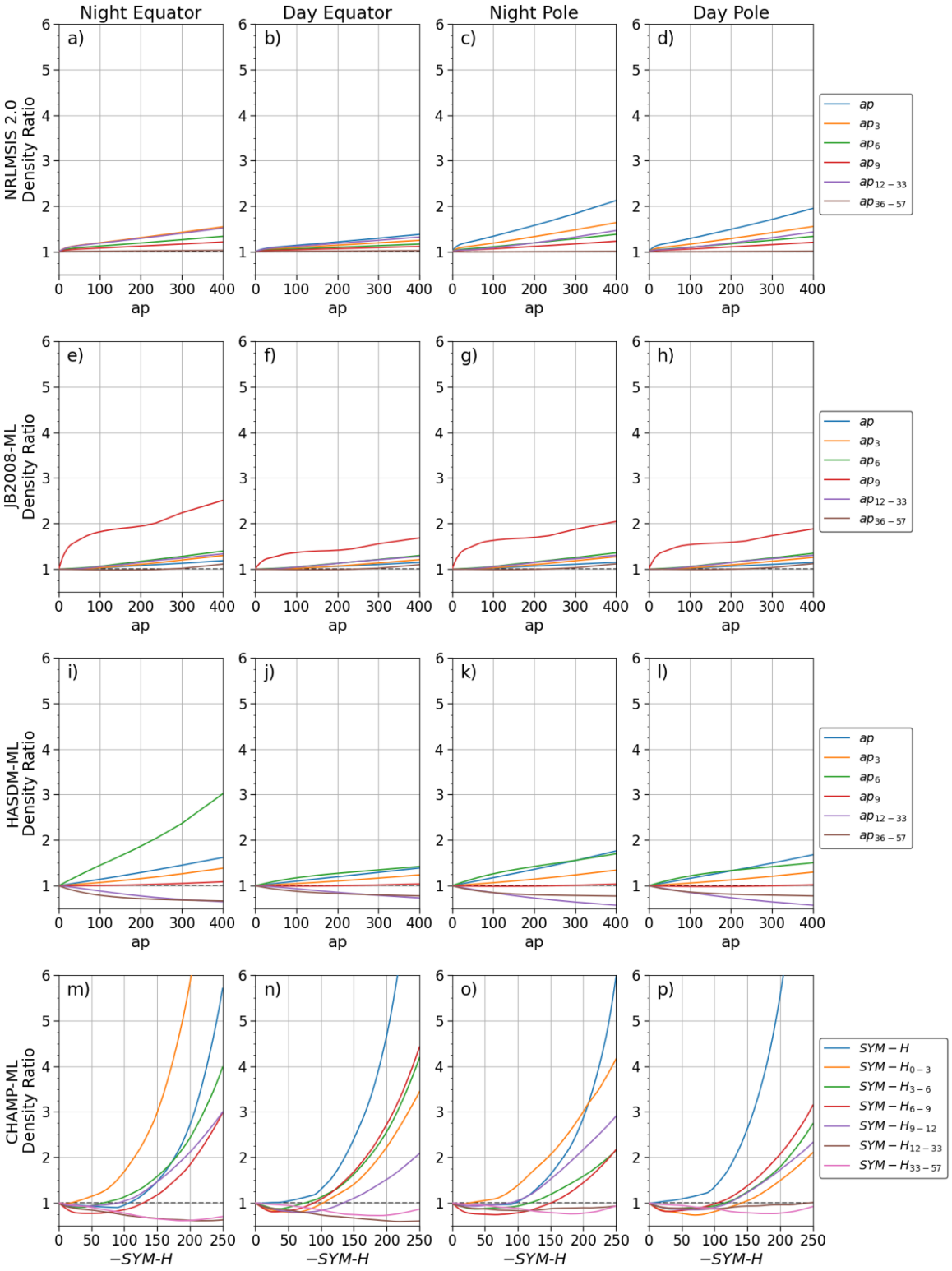

Figure 5.2: Density ratios for the four models and four locations. The ratios are computed with respect to the particular driver being set to zero.

While there are values for JB2008-ML in Figure 5.2 that are less than 1.00, they are at most showing a 2% decrease and only at the equatorial locations.

HASDM-ML has a near-linear relationship with *ap*, but there is considerable evidence of post-storm cooling seen in Figure 5.2. At each of the four locations, $ap_{12\text{-}33}$ and $ap_{36\text{-}57}$ have an inverse relationship with density. At the two high-latitude locations, $ap_{12\text{-}33}$ causes the lowest density ratios while $ap_{36\text{-}57}$ causes the lowest density ratios at the equator. This may be a result of the time-delay of the density response at low latitudes relative to the auroral region. In contrast to JB2008-ML panels (e)-(h), HASDM-ML has a strong positive relationship between ap_6 and density with little impact from ap_9. At the highest levels of activity ($ap > 300$), the current *ap* value drives the strongest increase in density at the poles.

CHAMP-ML displays a highly nonlinear relationship between *SYM-H* and density. At each location, the relative importance of each input can change significantly; the maximum density ratio for $SYM\text{-}H_{0\text{-}3}$ is 10.25 at the nightside equator while it is only 2.10 at the dayside pole. There is strong evidence of post-storm cooling in the CHAMP dataset, highlighted by the array of historical *SYM-H* drivers causing density ratios below 1.00. The least recent *SYM-H* averages have their strongest inverse relationship with density at the equatorial locations while other historical indices demonstrate low density ratios at the polar location. The CHAMP-ML density ratios drop as low as 0.59 and rise as high as 12.34 indicating a more complex relationship between geomagnetic activity and density compared to the other three models in this analysis.

5.3.1 MSIS-UQ Cooling Study

As we had done in the previous section, we show the orbit-average density along CHAMP's orbit during the 2003 Halloween storm for CHAMP, NRLMSIS 2.0, EXTEMPLAR, and MSIS-UQ in Figure 5.3. EXTEMPLAR is included as it is a linear approach to the exospheric temperature modeling for NRLMSIS 2.0.

In the pre-storm period (10/28–10/29), there is again variability in model outputs. During the first peak of the storm, the models show similar trends but are mostly above the CHAMP density estimates. EXTEMPLAR overpredicts density here more than the other two models. In the lull

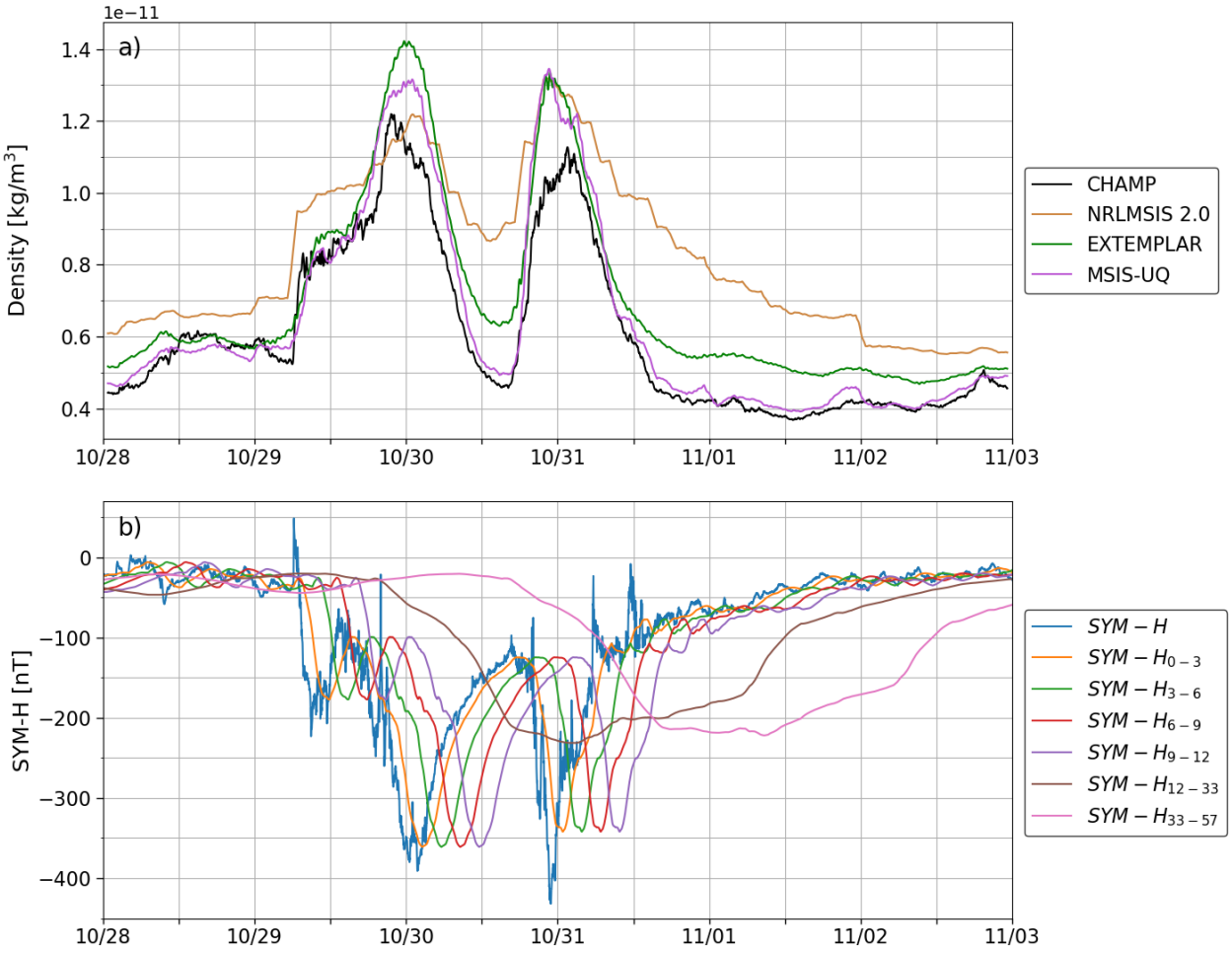

Figure 5.3: Orbit-average density for NRLMSIS 2.0, EXTEMPLAR, MSIS-UQ, and CHAMP (a) and the associated *SYM-H* time-series inputs (b).

between the two peaks (10/30–10/31), all models show density decreases but to varying extents. NRLMSIS 2.0 shows the smallest density decay during this period, but due to the exospheric temperature corrections in MSIS-UQ, the density falls to CHAMP levels. EXTEMPLAR still significantly reduces density in this period. For the second storm, all models overestimate density again to very similar extents. In the post-storm period (10/31–11/03), NRLMSIS 2.0 does not follow the trend observed by CHAMP, as was noted in Figure 5.1. EXTEMPLAR significantly improves upon NRLMSIS 2.0, but MSIS-UQ follows the post-storm overcooling trends observed by CHAMP with little bias. This highlights the impact the exospheric temperature corrections to

NRLMSIS 2.0 have. We now conduct the overcooling study once more but focus on the impact MSIS-UQ has on NRLMSIS 2.0. These results are shown in Figure 5.4.

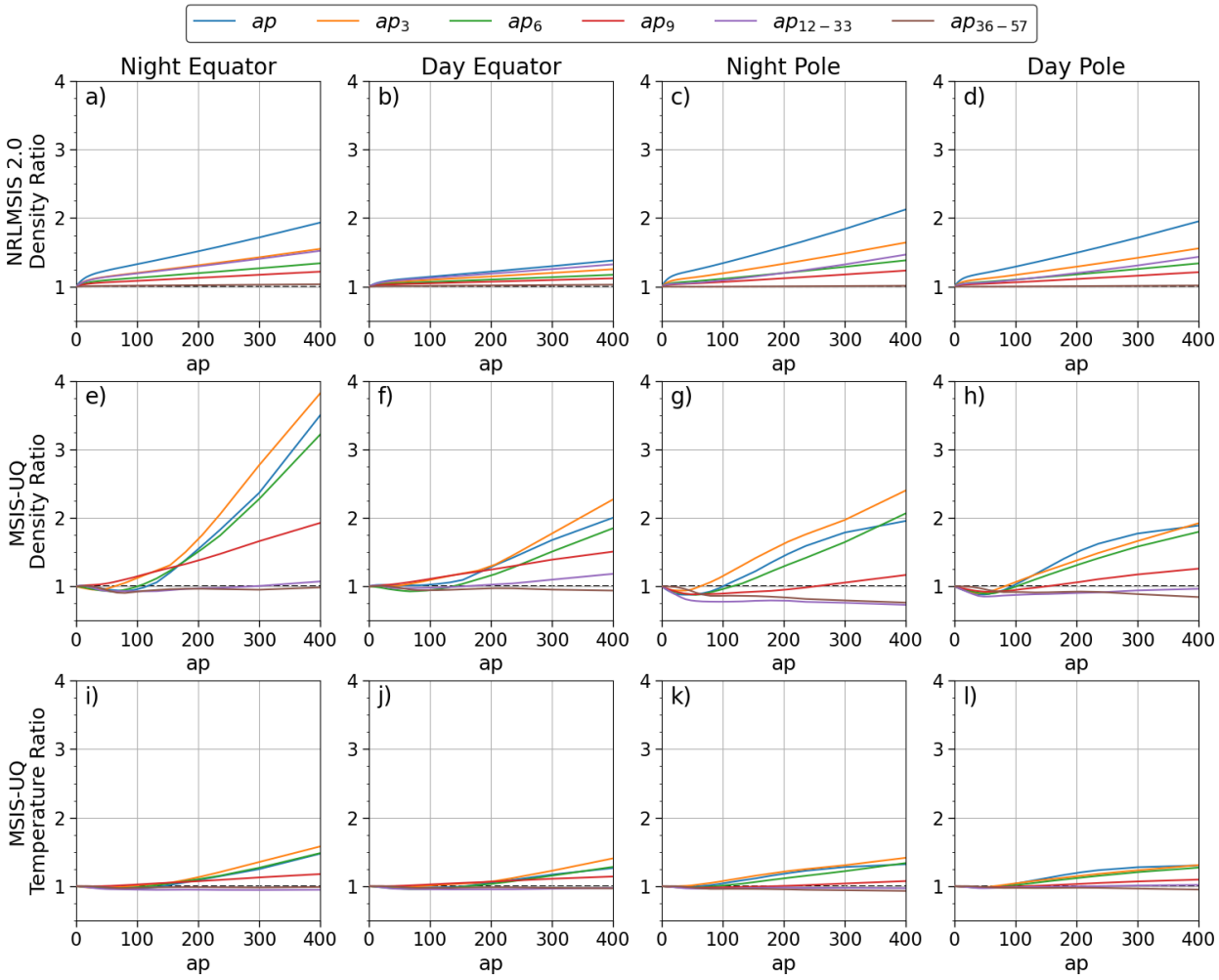

Figure 5.4: Density ratios for NRLMSIS 2.0 (a–d) and MSIS-UQ (e–h) with the corresponding temperature ratios for MSIS-UQ (i–l).

Panels (a)–(d) in Figure 5.4 show the density ratios for NRLMSIS 2.0 while increasing each time-series *ap* value independently. These results were discussed for Figure 5.2. The second row (panels (e)–(h)) shows the results for MSIS-UQ. The trends shown in these panels contradict many observations when using NRLMSIS 2.0 alone. For example, at 3/4 locations, ap_3 causes the largest density ratios, even being nearly twice as large at the night equator. MSIS-UQ also

shows a nonlinear relationship between geomagnetic activity and density which is not as clearly seen in NRLMSIS 2.0 alone. MSIS-UQ also enforces the idea of negative density ratios – density decreasing while the geomagnetic drivers are increasing. This is most pronounced at the two pole locations. When the least recent geomagnetic drivers ($ap_{12\text{-}33}$ and $ap_{36\text{-}57}$) become large, the density becomes up to 25% lower than when they are set to zero. This overcooling was seen in Figure 5.3 and is observed in the satellite density data, therefore becoming present in MSIS-UQ.

Another interesting trend is seen at low levels on geomagnetic activity particularly in panels (g) and (h). When any of the *ap* values increase from 0 up to 50–100, the density decreases. This seems counter-intuitive but could be caused by the approach of the study. When *ap* is being considered, for example, the *ap* and *SYM-H* values are set to 56 and -50 *nT*, respectively. When $ap = 0$, this would represent the time immediately after moderate geomagnetic activity while $ap = 56$ would represent sustained moderate geomagnetic activity. The model shows that when the conditions abruptly return to quiet values, the density increases – likely only temporarily. The bottom row (panels (i)–(l)) show the temperature ratios from MSIS-UQ corresponding to the middle row. The general trends are the same between temperature and density; however, the difference comes from the magnitude. The relative changes in temperature result in much stronger changes in density. There are negative temperature ratios, but they are much less prominent due to the consistent scaling.

5.4 Summary

This chapter diverts from an operational focus to study the scientific value of ML in space weather and the thermosphere. We demonstrated the use of machine-learned models to quantify the behavior of thermospheric post-storm cooling. We used three of the models from Chapter 4 in addition to JB2008-ML to conduct this assessment and used NRLMSIS 2.0 for comparison. All models were provided a recent time history (up to 57 hours) of geomagnetic drivers to see if the data suggests that there is evidence of post-storm cooling; the models would need to see that previous geomagnetic drivers indicate a storm of a given strength has recently occurred. Using the 2003 Halloween storms as an example, we showed that both NRLMSIS 2.0 and JB2008-ML do not

match the sudden cooling seen between and after the two storms by the CHAMP accelerometer. Meanwhile, HASDM-ML, CHAMP-ML, and MSIS-UQ all model the general density trends of this storm and display attributes of an abruptly cooled thermosphere.

When considering a historical event, other factors play a role in how the thermosphere behaves. Therefore, we isolated the internal model formulation only as it pertains to recent magnetic perturbations. We held all model drivers constant and only varied a single geomagnetic driver at a time: *ap* for NRLMSIS 2.0, JB2008-ML and HASDM-ML, *SYM-H* for CHAMP-ML, and both for MSIS-UQ. This showed that NRLMSIS 2.0 and JB2008-ML both did not exhibit any cooling effects as the historical *ap* values were raised, which would indicate a strong storm had recently taken place. Conversely, HASDM-ML, CHAMP-ML, and MSIS-UQ all showed evidence of strong post-storm cooling with density ratios as low as 57%–75% of the baseline magnitude for the least recent drivers. Showing this model's ability to portray this phenomenon is important, as this is an area where many models struggle (e.g. NRLMSIS 2.0). Following an extreme event, a strong overestimation of density could lead to significant error in a satellite's predicted state, potentially wreaking havoc in a conjunction assessment.

Chapter 6. Benchmarking Space Weather Driver Forecasting Models

All of the evaluations thus far were focused on errors in density modeling. However, errors in space weather driver forecasts cause errors in the resulting densities, therefore impairing satellite conjunction analyses. Bussy-Virat et al. [135] performed a study to show the effects of driver uncertainty on the probability of collision between two space objects. In order to achieve this, the authors performed an analysis on two years of F_{10} and ap forecast errors. This is expanded upon here by using (i) all solar and geomagnetic drivers that are used in operations, (ii) a large historical data set covering a period of six (6) years, (iii) an extended forecast window of up to six (6) days, and (iv) the initial driver values to characterize model performance as a function of the solar and geomagnetic activity. This expansion is performed to get a more complete picture of the legacy drivers and specifically analyze the performance of the driver forecasts that are directly fed to JB2008 and subsequently HASDM.

6.1 Methodology

The proprietary SET algorithms automatically produce files every three hours generating updated six-day forecasts for solar and geomagnetic indices and proxies. The forecasts are used with exclusive, restricted access by the USAF customer. This study is the first time that the metrics of the forecasts have been evaluated and made public. These forecasts have a temporal resolution of three hours. In addition, they archive the observed values for each time step. To conduct this analysis, forecasts from October 2012 through the end of 2018 were used with the exception of a small number of missing/corrupted forecasts. In total, there were over 15,000 files to leverage for this study.

In order to effectively examine the solar and geomagnetic indices in comparable terms, a consistent approach had to be determined. To provide the clearest possible representation for all indices, different methods are used for solar indices/proxies and geomagnetic indices but kept consistent within each of the domains. Each index was split into separate sub-populations depending

on the forecasted values. Populations that ended up with fewer than 100 forecasts are not shown, because there is insufficient data to draw statistical conclusions.

6.1.1 Solar Indices

The task of generating statistical results for the four solar indices investigated (F_{10}, S_{10}, M_{10}, and Y_{10}) is relatively straightforward. The forecasts are generated using SET's *SOLAR2000* algorithm [136, 137]. The thresholds to assess activity level for F_{10} and *ap* have been described by [105, 138] and are combined here with a supplementary statistical analysis for the remaining solar indices and proxies. The objective in setting thresholds is to group data by general solar activity levels. Figure 6.1 depicts how the solar indices are distributed based on the initially forecasted value (one day from forecast epoch), and Table 6.1 describes the solar activity levels.

Using these partitions on the 15,000+ forecasts resulted in a distinct number of individual F_{10} forecasts for each activity level. These were used to classify the remaining solar indices and proxies, with the absence of a natural partition, or lull, in the distribution. A natural partition for S_{10} can be seen at 150 sfu (panel (b)). This was chosen for that particular threshold as it did not greatly disrupt the number of forecasts in the adjacent activity levels since the goal was to have a similar number of forecasts across all solar indices and proxies for a given activity level. Peaks in the Figure 6.1 distribution data are a result of the natural distributions of solar activity estimated in a 3-hour cadence. Reading from right to left in the figures (high to low solar activity), the decline of solar cycle 24 from 2012 to 2018 is clearly portrayed and is the source of the predominantly bi-modal distributions.

Figure 6.1 shows how the forecasts are distributed and that all activity levels have sufficient data to perform the following analyses. Note that the shapes of the distributions within each activity level are not indicative of the distributions of the forecast errors within them. The four levels of solar activity are defined in Table 6.1.

With each index's/proxy's forecast appropriately divided on its initial forecasted value, uncertainty distributions could be generated with respect to time from the forecast epoch. The uncertainty for the solar indices is defined as the error with respect to the issued (actual archival) value,

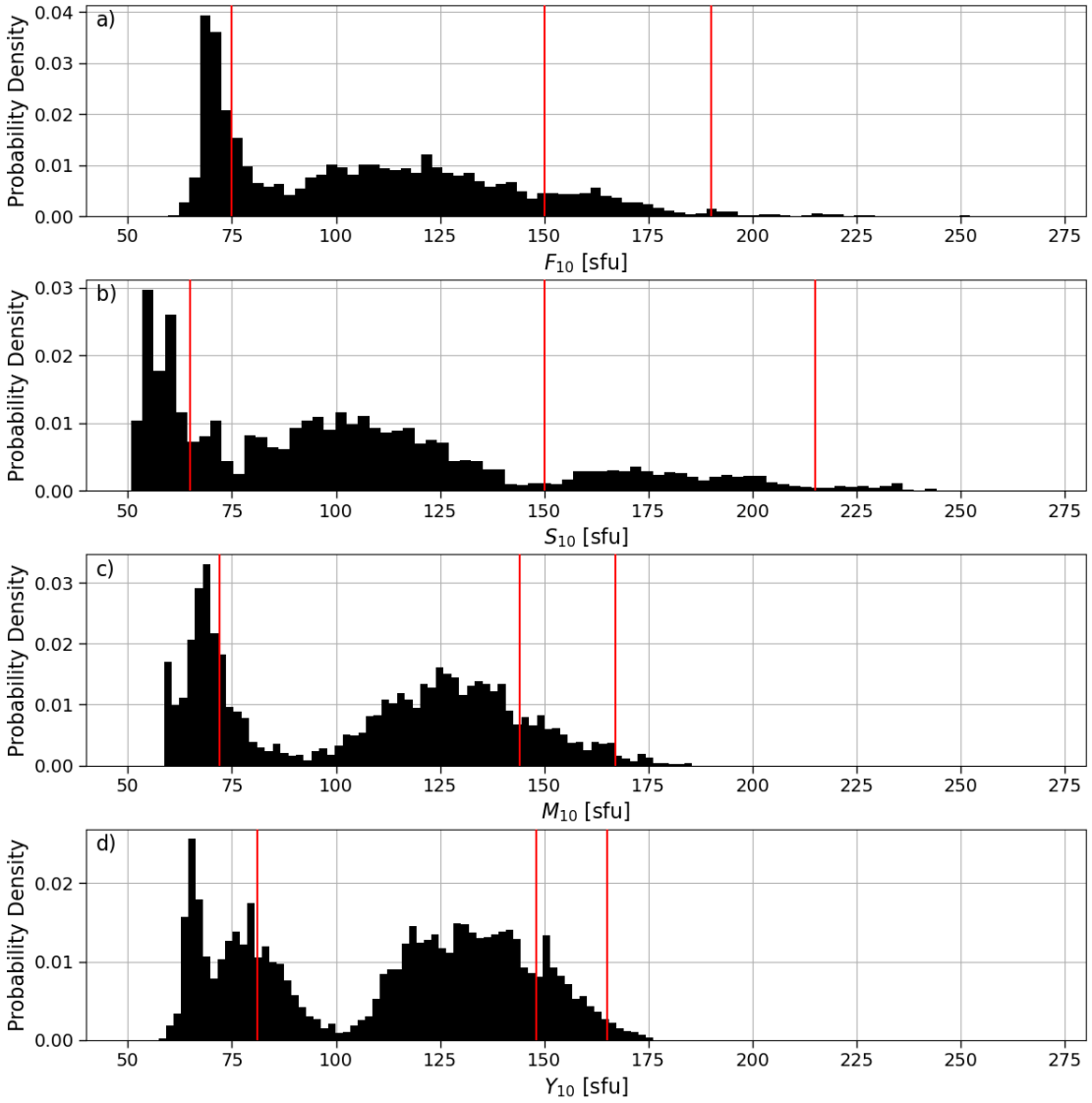

Figure 6.1: Distributions of initially forecasted values for each solar index with partitions shown in red.

normalized by the issued value. It is important to note that all errors shown (for both solar and geomagnetic indices) have a consistent sign convention. Positive percentages represent a forecasted value that was **more positive** than the issued (actual) value. For the solar indices and proxies, the error in solar flux units is also provided. All of the solar indices are updated daily, so there are twenty-four distributions for each (four magnitude-based and six temporal partitions).

Table 6.1: Activity level thresholds and units for the four solar indices.

	Activity Level			
Solar Driver	**Low**	**Moderate**	**Elevated**	**High**
F_{10} **[sfu]**	$F_{10} \leq 75$	$75 < F_{10} \leq 150$	$150 < F_{10} \leq 190$	$F_{10} > 190$
S_{10} **[sfu]**	$S_{10} \leq 65$	$65 < S_{10} \leq 150$	$150 < S_{10} \leq 215$	$S_{10} > 215$
M_{10} **[sfu]**	$M_{10} \leq 72$	$72 < M_{10} \leq 144$	$144 < M_{10} \leq 167$	$M_{10} > 167$
Y_{10} **[sfu]**	$Y_{10} \leq 81$	$81 < Y_{10} \leq 148$	$148 < Y_{10} \leq 165$	$Y_{10} > 165$

6.1.2 Geomagnetic Indices

The analysis of the two geomagnetic indices, *ap* and *Dst*, is more intricate. Not only are the uncertainties functions of their magnitudes and time from epoch, they vary with solar activity level. To analyze *ap*, three geomagnetic activity levels were chosen: low, moderate and active. In analyzing *Dst*, six geomagnetic activity levels were chosen and are consistent with the NOAA G-scale as operationally applied by SET. To allocate the geomagnetic forecasts, the largest value in the forecast for *ap* and the most negative value for *Dst* are the controlling factors. Figure 6.2 shows how the two geomagnetic indices are distributed based on these characteristics.

The *ap* distribution shows strong decay in forecast frequency with increasing *ap* values. There is a noticeable number of forecasts with a maximum value of 50 which get classified as moderate geomagnetic activity, even though the histogram shows the bar in the next activity level. The *Dst* distribution had such a significant amount of minimum forecasted values at or near zero that the distribution is also shown in a log-scale. This increases the visibility of areas with fewer forecasts. As previously noted, these distributions are not indicative of the forecast error distributions. Table 6.2 explicitly states the thresholds for *ap* and *Dst*.

In addition to the geomagnetic conditions, the forecast is classified by the initial forecasted F_{10} value. Since the distributions have a finer temporal resolution and a solar dependency, there are 576 and 1,152 distributions for *ap* and *Dst*, respectively. The geomagnetic indices could be classified by other metrics instead of F_{10}, such as average thermospheric density, but F_{10} was used since that was the most readily available additional metric.

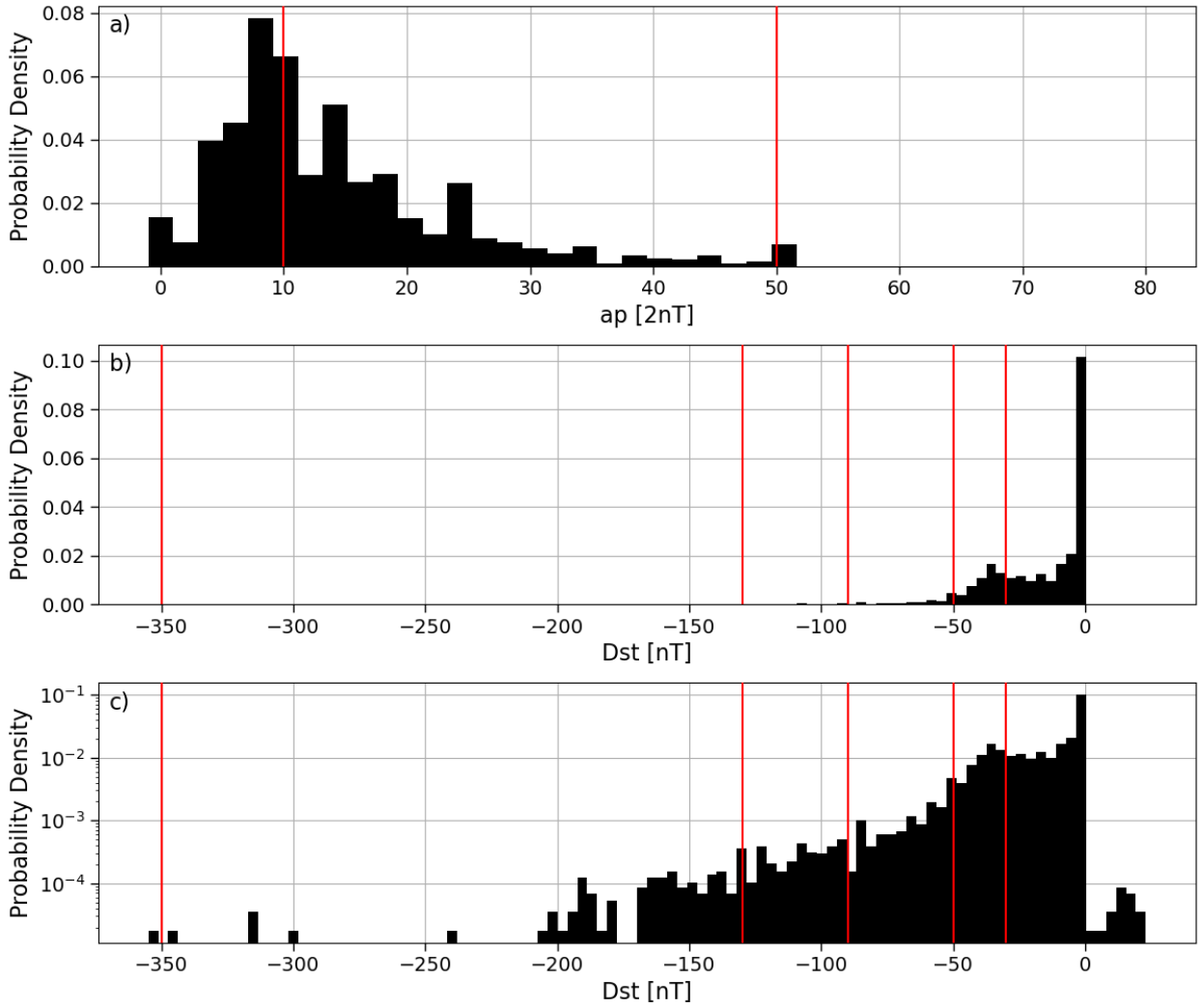

Figure 6.2: Distributions of initially forecasted values for the two geomagnetic indices with partitions shown in red. The *Dst* distribution is shown a second time with the frequency on a logarithmic scale for improved reading.

It becomes difficult to generate a standard percent error normalized by the issued value, because the issued value can be small or even zero. Therefore, no normalized errors are shown. The statistics provided in the proceeding section are the mean, standard deviation, and the error bound for the population mean (EBM). These are generated only for the forecast errors in the proxy's/index's units. Equations 6.1 and 6.2 show how the errors are computed in both absolute

Table 6.2: Activity level thresholds and units for geomagnetic activity, *ap* and *Dst*.

Index	Activity Level	Index Range
***ap* [2nT]**	**Low**	$ap \leq 10$
	Moderate	$10 < ap \leq 50$
	Active	$ap > 50$
***Dst* [nT]**	**G0**	$Dst \geq -30$
	G1	$-30 > Dst \geq -50$
	G2	$-50 > Dst \geq -90$
	G3	$-90 > Dst \geq -130$
	G4	$-130 > Dst \geq -350$
	G5	$Dst \leq -350$

terms and in percentage form.

$$\text{Error} = forecast - issued \tag{6.1}$$

$$\text{Percent Error} = 100\% \cdot \frac{forecast - issued}{issued} \tag{6.2}$$

In order to account for the sample mean not perfectly representing the population mean, the 95% confidence EBM is provided which can be used to determine the 95% confidence interval for the population mean. This is shown in Equations 6.3 and 6.4. $\text{CI}_{95\%}$, $\bar{x}$, σ, and n represent the 95% confidence bounds, the sample mean, the standard deviation, and the number of samples, respectively.

$$EBM = Z_{95\%}\frac{\sigma}{\sqrt{n}} \tag{6.3}$$

$$\text{CI}_{95\%} = \bar{x} \pm EBM \tag{6.4}$$

The Z value that corresponds to 95% confidence is 1.9600, and in the proceeding tables, the EBM values are given with respect to the standard deviation in the respective units, not the normalized form. Other EBMs and confidence intervals can be easily computed using the corresponding Z

value and the values in Tables 6.3-6.8.

6.2 Results

In the resulting uncertainty figures, the mean and standard deviation of forecast error (as a function of time from forecast epoch) are presented for each activity level. This way, biases can be identified and the algorithm's temporal uncertainty can be determined. Figure 6.3 shows the performance of the F_{10} forecast algorithm, and Table 6.3 shows the statistics in sfu.

Table 6.3: Distribution statistics F_{10} error distributions (Figure 6.3).

Condition	Statistics	1 Day	2 Days	3 Days	4 Days	5 Days	6 Days
Low Solar	μ	-0.2685	-0.7472	-0.6672	-0.3721	-0.0674	0.2428
	σ	3.6985	4.7031	5.5001	6.1683	6.7677	7.2050
	EBM	0.1126	0.1432	0.1675	0.1878	0.2061	0.2194
Moderate Solar	μ	-0.8251	-0.8095	-0.9639	-1.1450	-1.1456	-1.1679
	σ	12.0854	14.9853	17.8425	20.2973	21.9353	23.3389
	EBM	0.2489	0.3086	0.3674	0.4180	0.4517	0.4806
Elevated Solar	μ	5.7270	7.2425	9.0385	10.3829	11.0017	10.9559
	σ	18.3328	22.1021	25.2942	27.0774	27.5279	26.9074
	EBM	0.8572	1.0335	1.1827	1.2661	1.2872	1.2582
High Solar	μ	15.7448	19.5749	24.2444	26.6674	26.0230	23.9778
	σ	20.2227	24.7236	27.6092	30.8869	33.9069	35.9795
	EBM	2.5639	3.1345	3.5003	3.9159	4.2988	4.5616

At low and moderate levels of solar activity, the F_{10} algorithm is fairly unbiased. It is not until elevated and high solar activity that a bias accumulates, showing a tendency of over-forecasting the proxy. The evolution of the error's standard deviation has an expected growth with time from epoch for all activity levels, showing the uncertainty of the forecast increasing with time. The algorithm performs well when the first forecasted F_{10} value is below 150 sfu, which accounted for approximately 87% of the forecasts. This analysis points to needed improvements in F_{10} prediction for periods of elevated and high solar activity. For moderate solar activity, the normalized error has a slightly positive bias where the bias is slightly negative when looking at the actual mean errors

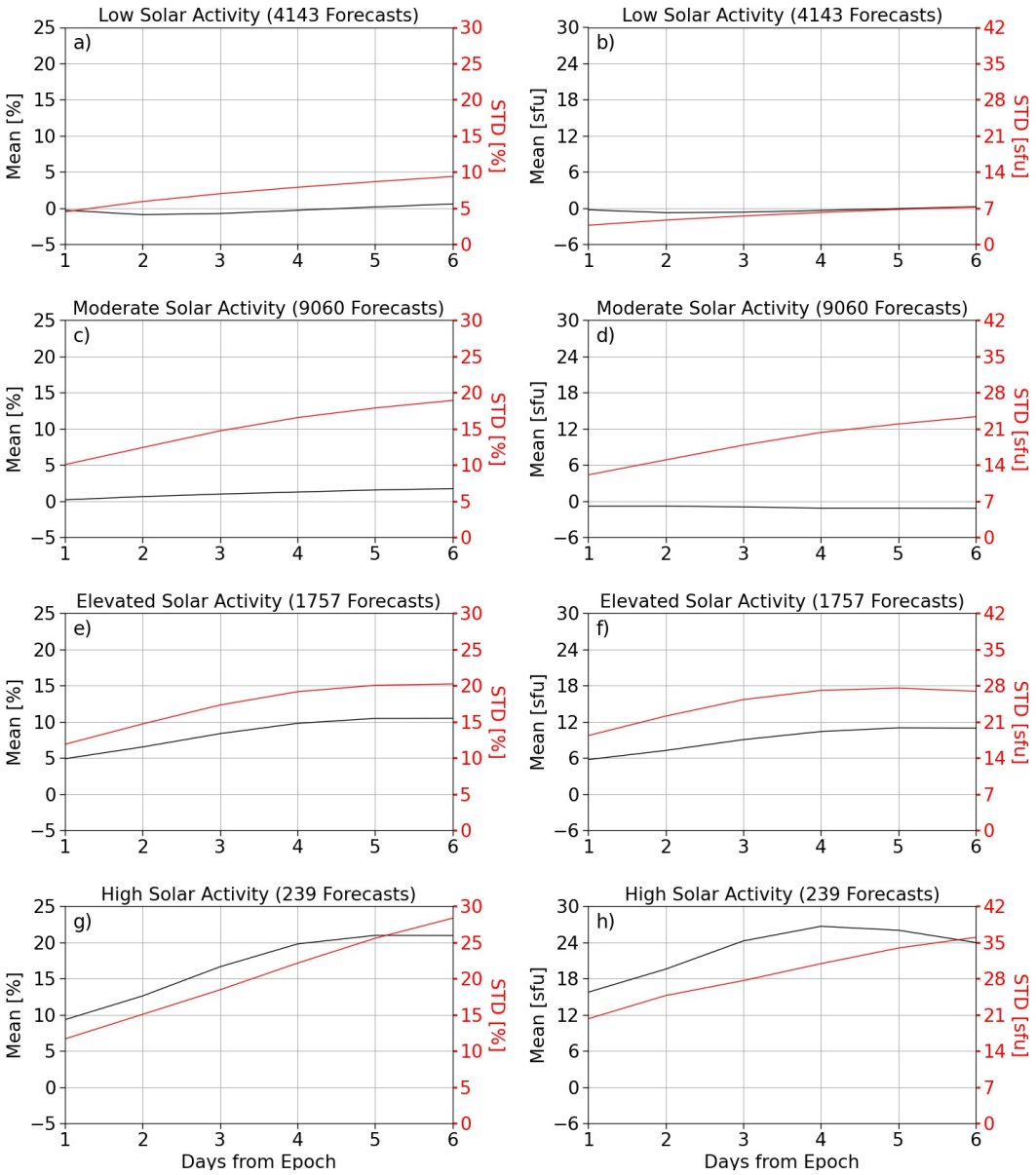

Figure 6.3: F_{10} algorithm performance across four levels of solar activity with normalized error shown on the left and absolute error shown on the right.

(to the right). This is caused by the range of this activity level (75 sfu to 150 sfu). This shows that the algorithm is likely over-forecasting F_{10} towards the higher end of the activity level and under-forecasting at the lower end. This would cause the normalized mean to rise relative to the actual mean errors. This is confirmed by panels (a,b) and (e,f) for low and elevated solar activity where the algorithm is under-forecasting and over-forecasting, respectively. This analysis on the

discrepancy between the normalized and actual mean errors is applicable to the remaining solar indices and proxies.

Figure 6.4 and Table 6.4 provides the algorithm performance for S_{10}. There is little bias through low, moderate, and elevated activity levels (over 98% of forecasts) displaying strong overall performance. The uncertainty at these activity levels is similar to F_{10}, but the performance at high solar activity is not as stable. For high solar activity, there is a dominant tendency to over-forecast in addition to a large uncertainty. This is a byproduct of the forecasting method. The uncertainty also does not consistently grow with time. Thus, S_{10} prediction needs more attention for high solar activity periods.

Table 6.4: Distribution statistics S_{10} error distributions (Figure 6.4).

Condition	**Statistics**	**1 Day**	**2 Days**	**3 Days**	**4 Days**	**5 Days**	**6 Days**
Low Solar	$\boldsymbol{\mu}$	0.2206	-0.0343	-0.1044	-0.0985	-0.0660	-0.0501
	$\boldsymbol{\sigma}$	5.8378	5.7474	5.8682	6.0709	6.1153	6.0475
	EBM	0.1797	0.1769	0.1806	0.1869	0.1883	0.1862
Moderate Solar	$\boldsymbol{\mu}$	0.1111	0.2190	0.2974	0.3426	0.2977	0.1862
	$\boldsymbol{\sigma}$	10.0715	11.7059	13.5434	14.9844	16.1132	17.1467
	EBM	0.2102	0.2443	0.2827	0.3127	0.3363	0.3579
Elevated Solar	$\boldsymbol{\mu}$	-1.1311	-1.7989	-2.2361	-2.6910	-2.8148	-2.7593
	$\boldsymbol{\sigma}$	16.5592	19.2050	21.9271	24.1118	25.8017	27.2658
	EBM	0.7120	0.8257	0.9428	1.0367	1.1094	1.1723
High Solar	$\boldsymbol{\mu}$	16.2040	23.1628	28.3943	31.4085	31.6158	30.1409
	$\boldsymbol{\sigma}$	35.3267	39.6577	39.6471	38.7875	38.5307	38.6613
	EBM	4.4060	4.9458	4.9445	4.8373	4.8052	4.8215

The F_{10} and S_{10} algorithms are both vulnerable to high solar activity, but the comprehensive effectiveness is visible. The limitation during high activity is due to the volatility of the Sun during solar maximum, i.e, the inability to accurately forecast flares and the lack of information from the solar East limb and solar far-side active region's growth. The algorithms for the remaining indices prove to be more robust to solar activity. The M_{10} performance is presented in Figure 6.5 and Table 6.5.

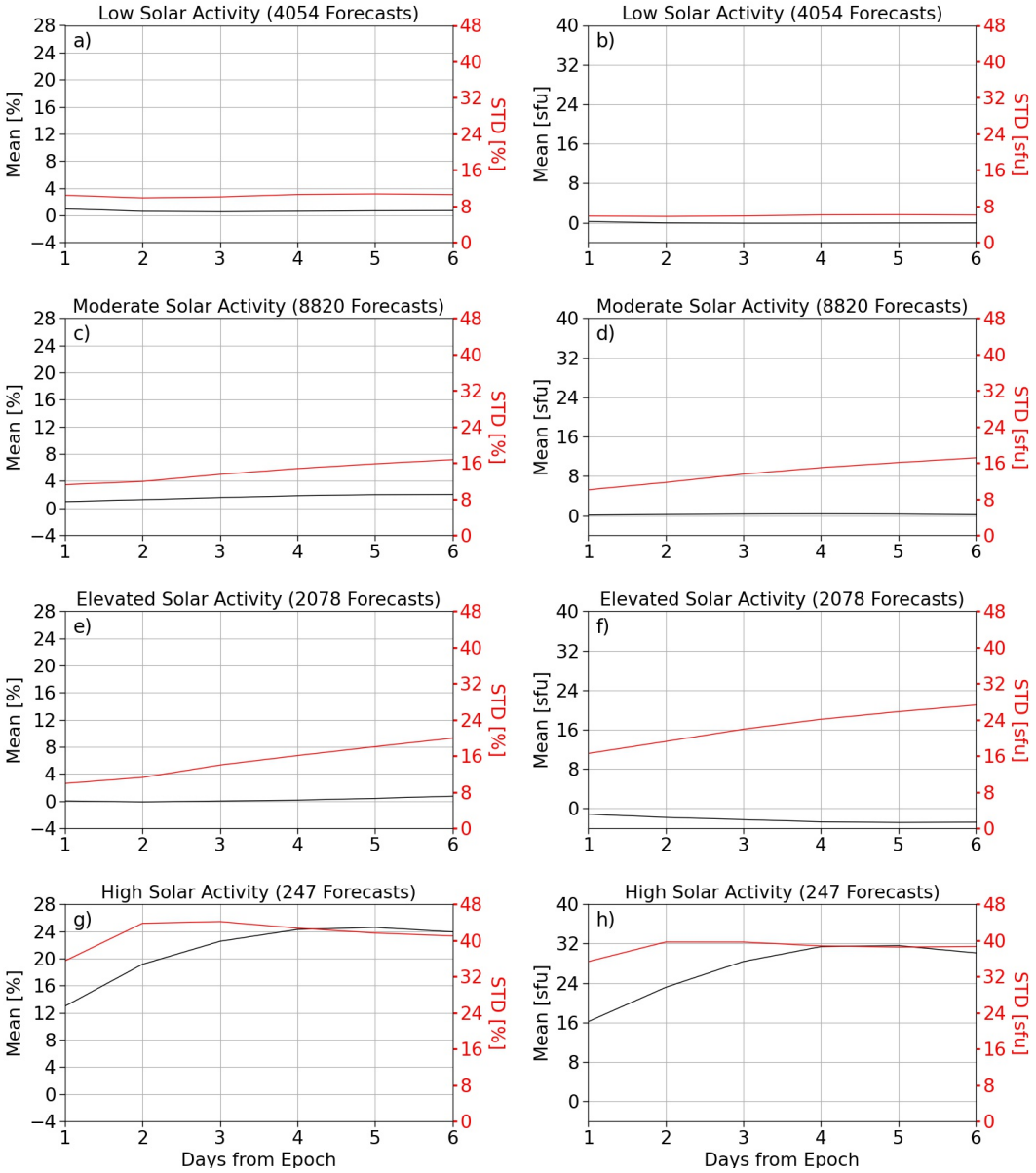

Figure 6.4: S_{10} algorithm performance across four levels of solar activity with normalized error shown on the left and absolute error shown on the right.

For M_{10}, there is a minimal bias of $\pm 2\%$ for the lower two activity levels, but panels (b) and (d) show that there is a slight tendency to under-predict. At elevated and high solar activity, the bias is accumulating with time and increases in intensity. Across all levels, the uncertainty starts below 4% and grows steadily with time. An interesting characteristic that contrasts the prior two indices is the lower uncertainty at high solar activity. The difference in performance is not drastic relative

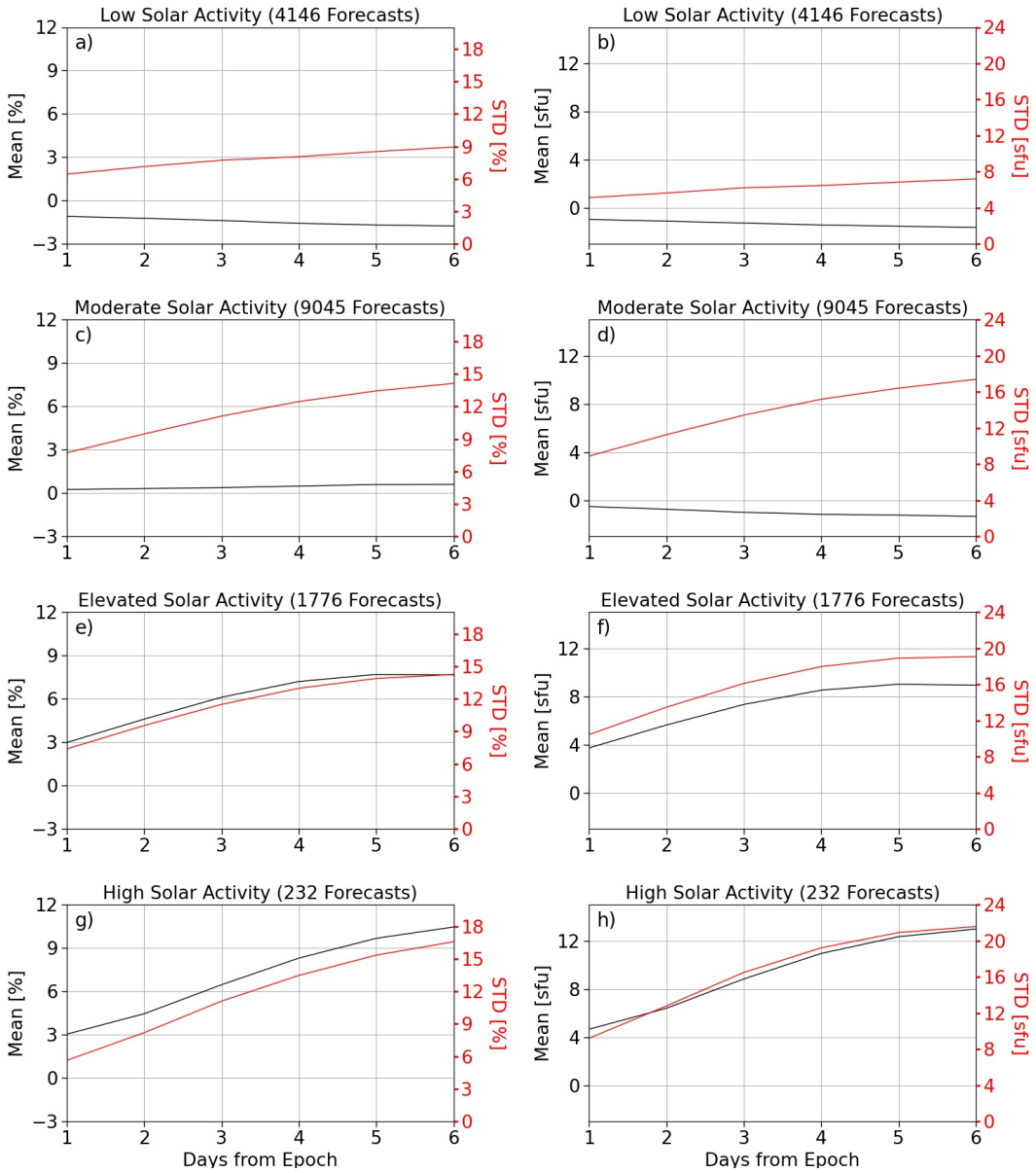

Figure 6.5: M_{10} algorithm performance across four levels of solar activity with normalized error shown on the left and absolute error shown on the right.

to the other conditions. Even so, improvement in M_{10} is needed for elevated and high solar activity periods.

To conclude the analysis of the solar indices, Figure 6.6 and Table 6.6 both show the performance of the Y_{10} algorithm. Relative to the previous three indices, the Y_{10} algorithm is considerably robust to activity levels and has less overall uncertainty. In the first two activity levels, the bias is

Table 6.5: Distribution statistics M_{10} error distributions (Figure 6.5).

Condition	Statistics	1 Day	2 Days	3 Days	4 Days	5 Days	6 Days
Low Solar	μ	-0.9582	-1.1063	-1.2667	-1.4203	-1.5303	-1.6171
	σ	5.1317	5.6375	6.2250	6.4693	6.8538	7.2024
	EBM	0.1562	0.1716	0.1895	0.1969	0.2086	0.2192
Moderate Solar	μ	-0.5138	-0.7424	-0.9976	-1.1619	-1.2370	-1.3348
	σ	8.9027	11.2745	13.4329	15.1962	16.4221	17.4041
	EBM	0.1835	0.2324	0.2768	0.3132	0.3384	0.3587
Elevated Solar	μ	3.7282	5.6301	7.3375	8.5258	9.0192	8.9334
	σ	10.4528	13.4784	16.1156	17.9805	18.9175	19.0770
	EBM	0.4861	0.6269	0.7495	0.8363	0.8798	0.8872
High Solar	μ	4.6517	6.3966	8.8328	10.9500	12.3435	12.9784
	σ	9.2079	12.7892	16.4932	19.2241	20.9130	21.5820
	EBM	1.1849	1.6457	2.1223	2.4738	2.6911	2.7772

less than $\pm 1\%$ for nearly the entire prediction window. The uncertainty grows with time for all activity levels, but its magnitude is less significant than the other indices. The bias never exceeds 5% and the uncertainty 12%.

Table 6.6: Distribution statistics Y_{10} error distributions (Figure 6.6).

Condition	Statistics	1 Day	2 Days	3 Days	4 Days	5 Days	6 Days
Low Solar	μ	-0.4270	-0.2282	-0.0684	0.0902	0.2697	0.4650
	σ	4.5093	4.9373	5.4416	5.9977	6.4255	6.8160
	EBM	0.1372	0.1502	0.1656	0.1825	0.1955	0.2074
Moderate Solar	μ	-0.8277	-1.3317	-1.7651	-2.1734	-2.5561	-2.9688
	σ	8.5137	10.1811	11.8475	13.0170	13.8737	14.8937
	EBM	0.1753	0.2096	0.2439	0.2680	0.2856	0.3066
Elevated Solar	μ	2.1089	1.7243	2.0679	2.5656	2.7714	2.8423
	σ	7.7173	9.3066	10.6086	11.4955	11.8996	11.9965
	EBM	0.3593	0.4333	0.4939	0.5352	0.5541	0.5586
High Solar	μ	5.2729	4.2075	4.5855	5.2131	5.6561	6.0140
	σ	8.5806	10.8097	11.5436	11.3048	11.0190	10.4416
	EBM	1.1497	1.4483	1.5466	1.5146	1.4761	1.3990

As previously stated, the geomagnetic indices were more difficult to analyze due to an increase

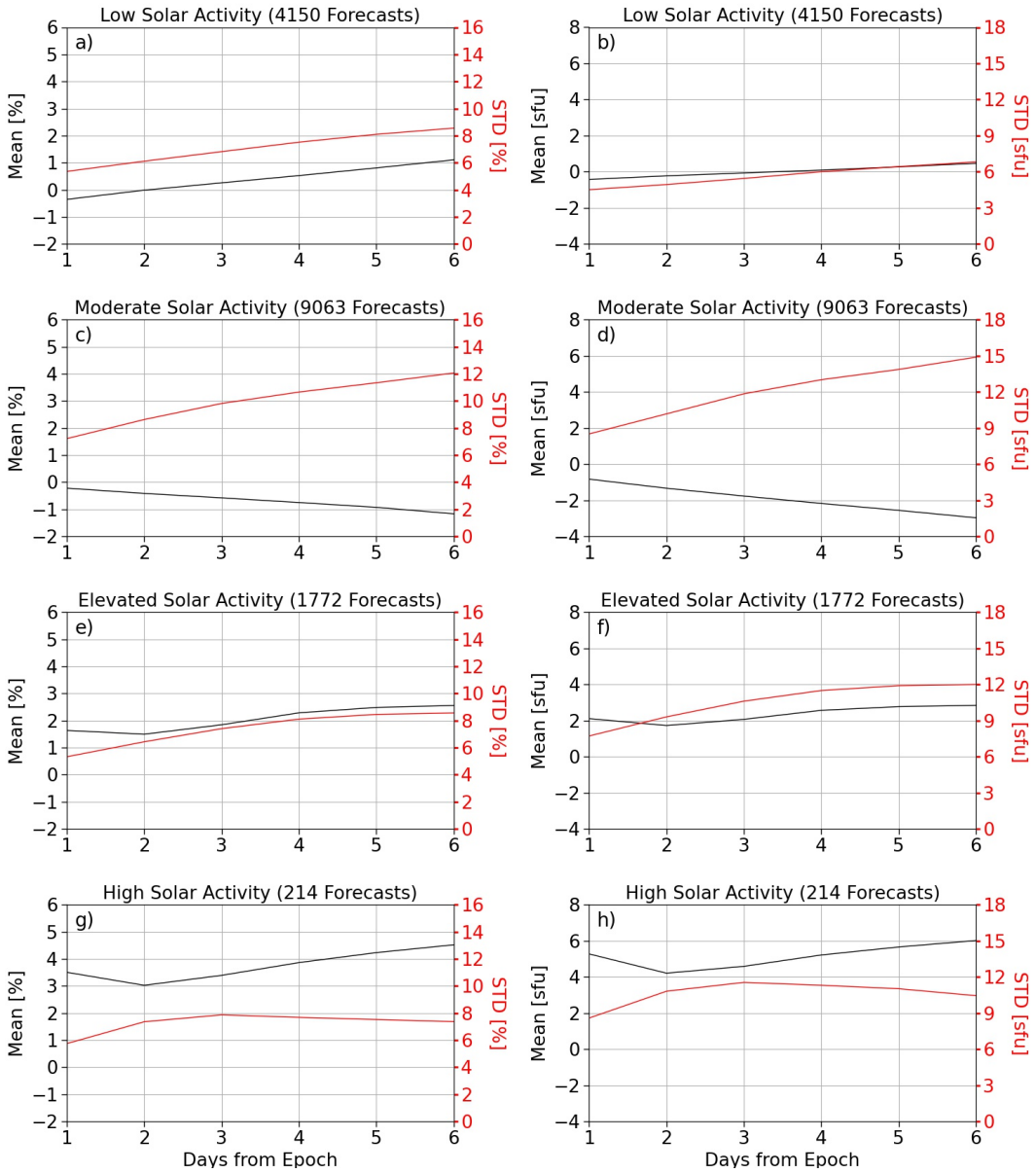

Figure 6.6: Y_{10} algorithm performance across four levels of solar activity with normalized error shown on the left and absolute error shown on the right.

in dependencies and a finer time resolution. Each geomagnetic index has its own set of activity levels but are both also based on F_{10} thresholds. The performance of the *ap* forecasts is shown in Figure 6.7 along with Table 6.7.

Unlike the solar indices, there are multiple conditions with insufficient data to conduct the analysis. The most distinct difference in the *ap* forecast performance, relative to the other indices,

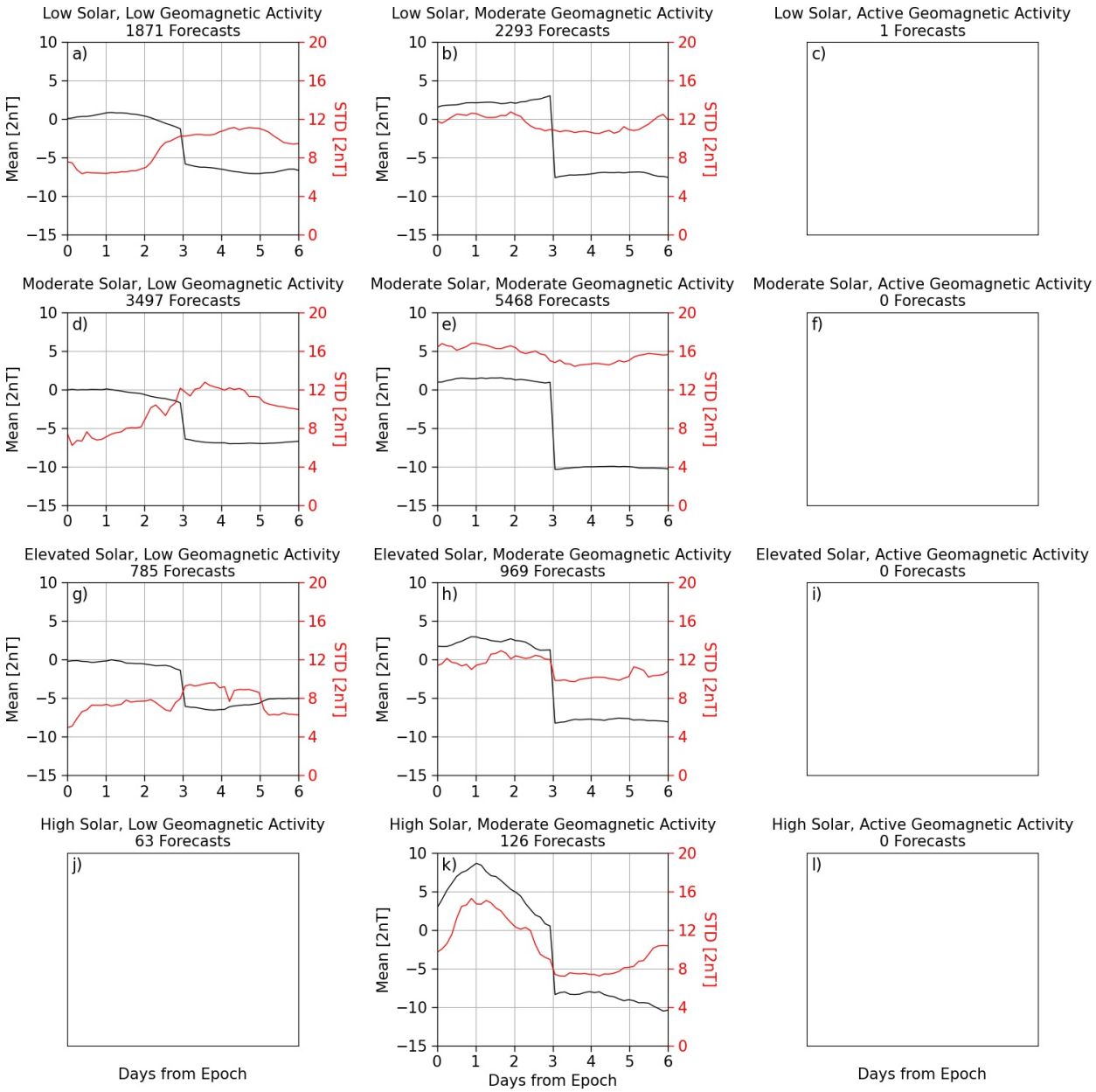

Figure 6.7: *ap* forecast uncertainty for the twelve solar and geomagnetic conditions in absolute terms.

is the discontinuity at the three-day mark. The forecasts only have a three-day prediction window. The forecasts are created by the ensemble of Space Weather forecasters at NOAA SWPC with the aid of the Geospace model.

Figure 6.7 shows uncertainty results for a six-day prediction window to be consistent with

Table 6.7: Distribution statistics for *ap* error distributions (Figure 6.7) in units of $2nT$. Days 1-3 represent the error statistics for the actual forecasts, where days 4-6 simply show background error that is a result of setting the forecast to zero.

Condition	Statistics	1 Day	2 Days	3 Days	4 Days	5 Days	6 Days
Low Solar Low Geomagnetic	μ	0.6782	0.4987	-1.2448	-6.4479	-7.0534	-6.6291
	σ	6.4053	6.8368	10.2313	10.6877	11.0401	9.4683
	EBM	0.2902	0.3098	0.4636	0.4843	0.5003	0.4290
Low Solar Moderate Geomagnetic	μ	2.1330	2.1653	3.0192	-7.1221	-6.8853	-7.5399
	σ	12.6072	12.7522	10.9242	10.6639	10.9207	11.9126
	EBM	0.5160	0.5220	0.4471	0.4365	0.4470	0.4876
Moderate Solar Low Geomagnetic	μ	-0.0492	-0.5465	-1.7532	-6.9294	-7.0114	-6.7315
	σ	6.8140	8.0992	12.1259	12.1145	11.2090	9.9162
	EBM	0.2258	0.2684	0.4019	0.4015	0.3715	0.3287
Moderate Solar Moderate Geomagnetic	μ	1.3877	1.3850	0.9151	-10.0225	-10.0000	-10.2871
	σ	16.7665	16.5156	14.9577	14.6091	14.9821	15.6026
	EBM	0.4444	0.4378	0.3965	0.3872	0.3971	0.4136
Elevated Solar Low Geomagnetic	μ	-0.2166	-0.5019	-1.3707	-6.4573	-5.6573	-5.0204
	σ	7.2675	7.6940	7.9520	9.0765	8.6062	6.2746
	EBM	0.5084	0.5382	0.5563	0.6350	0.6021	0.4389
Elevated Solar Moderate Geomagnetic	μ	2.9701	2.7038	1.2755	-7.7028	-7.6295	-8.0547
	σ	10.9766	12.0786	11.9995	10.0915	10.2440	10.7734
	EBM	0.6911	0.7605	0.7555	0.6354	0.6450	0.6783
High Solar Moderate Geomagnetic	μ	8.1667	5.2540	0.5079	-7.9921	-9.0397	-10.3968
	σ	15.2611	12.7671	8.9307	7.3996	8.1156	10.3549
	EBM	2.6647	2.2293	1.5594	1.2920	1.4171	1.8081

the other indices, even though SET sets every *ap* value to zero after three days. There are still interesting results in the latter three days of the forecasts across the different conditions. For example, the magnitude of under-prediction (when *ap* is set to zero) is different for each condition as is the volatility of *ap*, shown by the standard deviation. Even so, the most important aspect of two figures is the first three days when forecasts are provided, and this figure presents a benchmark on the prediction accuracy by NOAA SWPC for *ap*.

During low geomagnetic activity (across all solar activity levels), there is no significant bias detected. With moderate geomagnetic activity, there is a general over-prediction that decreases over the three-day provided forecast. It shows a possible path for prediction improvement by

relying on persistence when *ap* is high at the start of the forecasts. Another key determination is shown by the right-most panels where there is only a single forecast that has a value greater than 50. This reflects the difficulty in quantifying the intensity of a storm, even with the aid of a physics-based model.

The last algorithm analyzed is SET's *Anemomilos* for *Dst* forecasts, shown in Figure 6.8 along with Table 6.8. The G5 row is not shown since there was only a single forecast where a G5 storm was expected. There are only 9 of 24 conditions with enough forecasts to perform the analysis, but the remaining results provide insight to the strengths and weaknesses of the algorithm.

In panel (a) (when conditions are quiet), the forecasts remain relatively unbiased, and the uncertainty slowly increases with time. Figure 6.8 shows a general tendency to predict *Dst* to be more positive for nearly all G0 and G1 conditions, with the exception of G1 low solar activity conditions. In this case, the algorithm has a strong bias to expect *Dst* to be $\sim 23nT$ more negative than the issued values over the first four days of the forecast. Following the strong inclination after day four, the algorithm tends to neutralize the bias. This is interpreted as accurate prediction of *Dst* recovery to quiet conditions but over-prediction of the initial magnitude at the onset of the storm.

The bias for G1-G3, moderate solar activity conditions shows a strong temporal dependency transitioning from under to over prediction in each case. G2 moderate solar activity is a case with a peculiar trend of the uncertainty decaying with time from epoch. The algorithm tends to miss the magnitude and start of events and then achieves recovery to background after the main phase of the storm too quickly. This is also the case for G3 moderate solar activity, which shows even more pronounced error in this direction. This prediction error points to a need for improvement in understanding the arrival timing, magnitude, and duration of events. A source of the *Dst* over-prediction in G0-G3 conditions is that *Anemomilos* does not model (ignores) high-speed streams (HSS).

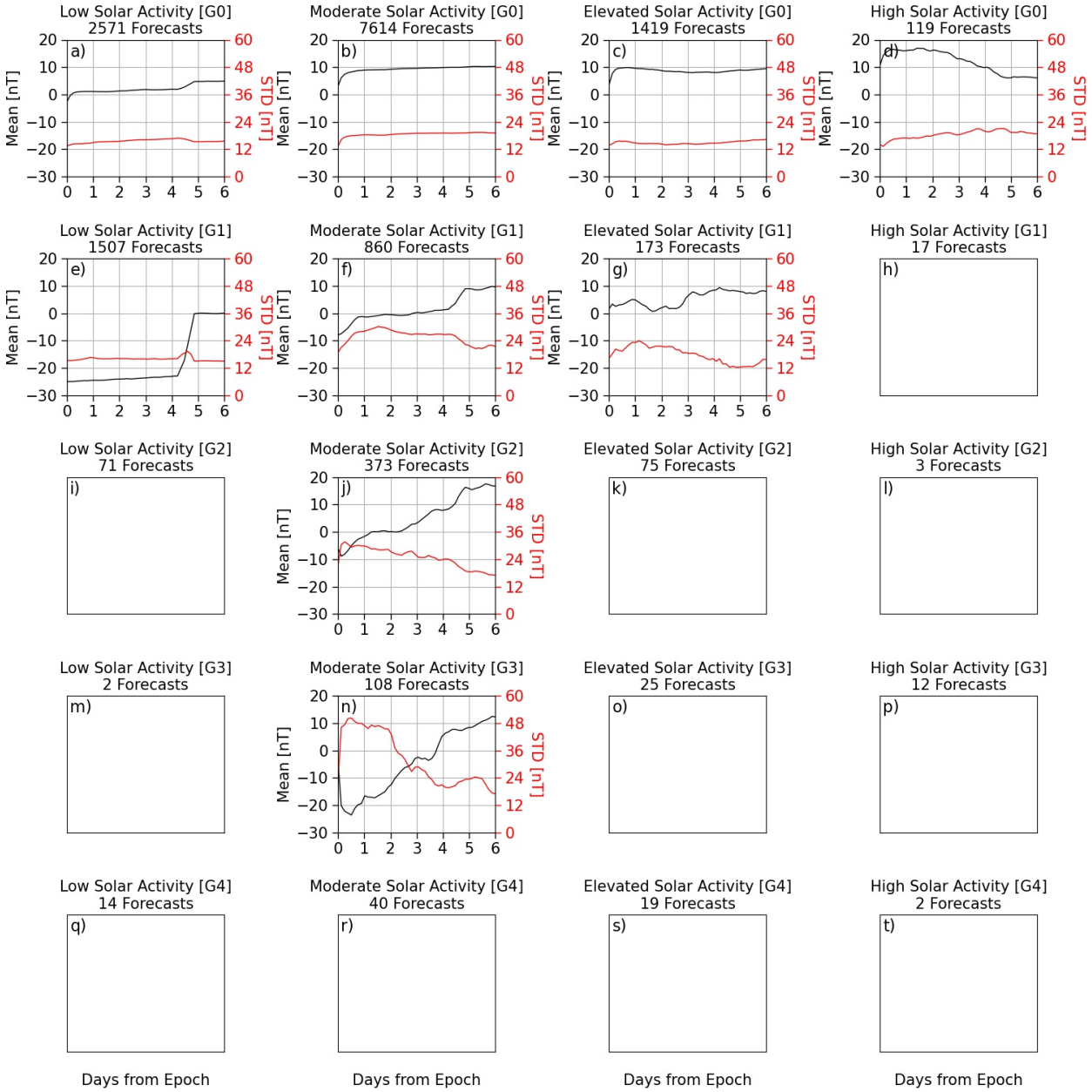

Figure 6.8: *Dst* forecast uncertainty for the combined solar and geomagnetic conditions in absolute terms.

Table 6.8: Distribution statistics for *Dst* error distributions (Figure 6.8) in units of nT.

Condition	Statistics	1 Day	2 Days	3 Days	4 Days	5 Days	6 Days
Low Solar G0	μ	1.1077	1.1851	1.8191	1.9479	4.7670	4.9067
	σ	14.6264	15.3676	16.0768	16.6027	15.2969	15.5632
	EBM	0.5654	0.5940	0.6214	0.6418	0.5913	0.6016
Moderate Solar G0	μ	8.8130	9.1455	9.5665	9.8593	10.1396	10.2912
	σ	18.2244	18.3000	18.9761	19.0591	19.2045	19.1569
	EBM	0.4094	0.4111	0.4262	0.4281	0.4314	0.4303
Elevated Solar G0	μ	9.6871	8.8767	8.0980	8.0500	8.9239	9.4235
	σ	14.8961	14.4229	14.4208	14.5799	15.5498	16.2407
	EBM	0.7751	0.7504	0.7503	0.7586	0.8091	0.8450
High Solar G0	μ	15.9664	15.9076	13.1176	9.8571	6.0840	6.0420
	σ	16.9190	17.6202	18.2560	20.0190	19.3606	18.7587
	EBM	3.0399	3.1659	3.2801	3.5969	3.4786	3.3704
Low Solar G1	μ	-24.6131	-24.1075	-23.7094	-23.1334	-0.0849	-0.0027
	σ	16.6175	16.1616	15.9475	15.9702	14.9486	14.9425
	EBM	0.8390	0.8160	0.8052	0.8063	0.7547	0.7544
Moderate Solar G1	μ	-1.2407	-0.5186	0.1395	1.1314	8.9733	9.5651
	σ	27.9324	29.1335	26.8674	26.7461	21.9025	21.3862
	EBM	1.8669	1.9471	1.7957	1.7876	1.4639	1.4294
Elevated Solar G1	μ	5.0405	1.6185	5.6069	8.3584	7.9191	7.9075
	σ	23.1907	21.5794	18.6853	16.0778	12.3314	15.6321
	EBM	3.4558	3.2157	2.7844	2.3959	1.8376	2.3294
Moderate Solar G2	μ	-2.1743	0.0804	2.8874	7.8525	15.9303	16.6944
	σ	29.8673	28.2948	26.2478	23.7456	18.4655	17.0010
	EBM	3.0311	2.8715	2.6638	2.4098	1.8740	1.7253
Moderate Solar G3	μ	-19.3704	-13.3611	-2.8704	5.1481	8.4444	12.3148
	σ	47.9640	45.4857	28.6925	20.9562	23.4494	16.9666
	EBM	9.0461	8.5787	5.4114	3.9524	4.4226	3.1999

6.3 Summary

This chapter focus on the errors for the current operational space weather driver forecasting models, providing a benchmark to all future algorithm/model development. We observed generally low forecasting errors for many of the drivers while there were certain conditions (e.g. high solar activity) that caused the error statistics to rise substantially. This would be an area in need of prompt improvement. As a community, we are continuously improving our capabilities, so it is important to have a benchmark. This can provide modelers a way to know when improvements have been made over the current operational standard. Another benefit to generating error statistic for these forecasting models is that they can be used to perturb the deterministic forecasts to investigate driver uncertainty in the absence of probabilistic driver forecasting models.

Chapter 7. Satellite Orbital Uncertainty Study

Up to now, the focus of this work has been on developing probabilistic thermosphere models (Chapter 4), examining their scientific value (Chapter 5), and extracting error statistics to benchmark the operational space weather driver forecasting models (Chapter 6). The work in Chapters 4 and 6 provide us with a unique opportunity to – for the first time – examine the effect of driver and model uncertainty on a satellite's state across diverse space weather conditions.

We consider the CHAMP satellite at an initial altitude of approximately 400 km. For reproducibility of this work, the initial state and satellite data for the ballistic coefficient are shown in Table 7.1. CHAMP was chosen for this study given its use in this dissertation and for its use in the field. The cross-sectional area corresponds to the satellite with zero-attitude, and the drag coefficient comes from models developed by Paul et al. [139]. The initial state and satellite parameters are used for each condition for consistency and comparison of the results. Only two-body, J_2 (effect of Earth's oblateness), and drag forces are considered.

Table 7.1: Initial state in the Cartesian reference frame and satellite parameters for conducting the analyses in this chapter.

X [m]	**Y [m]**	**Z [m]**
3782900.7032	-5441600.6779	-1420075.1327
V_X [m/s]	**V_Y [m/s]**	**V_Z [m/s]**
-606.6600	1539.2559	-7488.3946
C_D	**A [m^2]**	**m [kg]**
3.0912	0.7710	500

7.1 Driver Uncertainty

Seven time periods between October 2012 and December 2019 were chosen to perform this stdy, which coincides with the availability of operational deterministic forecasts from the previous chapter. The first four periods were chosen to cover each of the solar activity levels. For these

cases, the solar index samples were generated by perturbing the deterministic forecast using the temporal error statistics (mean and standard deviation at each point in time after forecast epoch) from Chapter 6. This is done 1,000 times for a Monte Carlo analysis, and the true geomagnetic driver variations are kept consistent between them. Table 7.2 shows the seven conditions with their respective start dates. Figure 7.1 shows the true, deterministic, and perturbed drivers for each of the four solar cases. Note: only F_{10} and *ap* are shown even though many of the models use additional inputs. This is discussed in Section 7.1.1.

Table 7.2: Activity levels and start dates for the seven periods considered in the satellite orbital uncertainty study.

Condition Name	**Activity Level**	**Start Date**
Solar 1	Low Solar	Nov. 14, 2018 (0900 UT)
Solar 2	Moderate Solar	Aug. 5, 2013 (0000 UT)
Solar 3	Elevated Solar	Jan. 13, 2013 (0900 UT)
Solar 4	High Solar	Oct. 27, 2014 (1800 UT)
Geo 1	Moderate Solar, Moderate Geomagnetic	Jun. 24, 2015 (2100 UT)
Geo 2	Elevated Solar, Low Geomagnetic	Mar. 3, 2014 (0000 UT)
Geo 3	High Solar, Moderate Geomagnetic	Jan 10, 2014 (1500 UT)

In the three-day periods considered in this study, F_{10} forecasts do not tend to significantly deviate from the true values with the exception of Solar 2 and the last day of Solar 4. However, models tend to be sensitive to solar activity [6]. In addition, perturbed forecasts are also generated for S_{10}, M_{10}, and Y_{10} based on their respective error distributions.

In Chapter 6, it had been observed that the *ap* forecast performance not only varied with geomagnetic activity level but also with solar activity level. Therefore, forecasts were distributed based on a combination of the two, resulting in twelve subpopulations for the temporal error statistics. For this study, only three conditions were chosen (see Table 7.2) to study the effect geomagnetic driver forecast uncertainty. For each of these conditions, the solar indices are kept at their true values. Meanwhile perturbed forecasts for *ap* and *Dst* are generated based on historical error statistics.

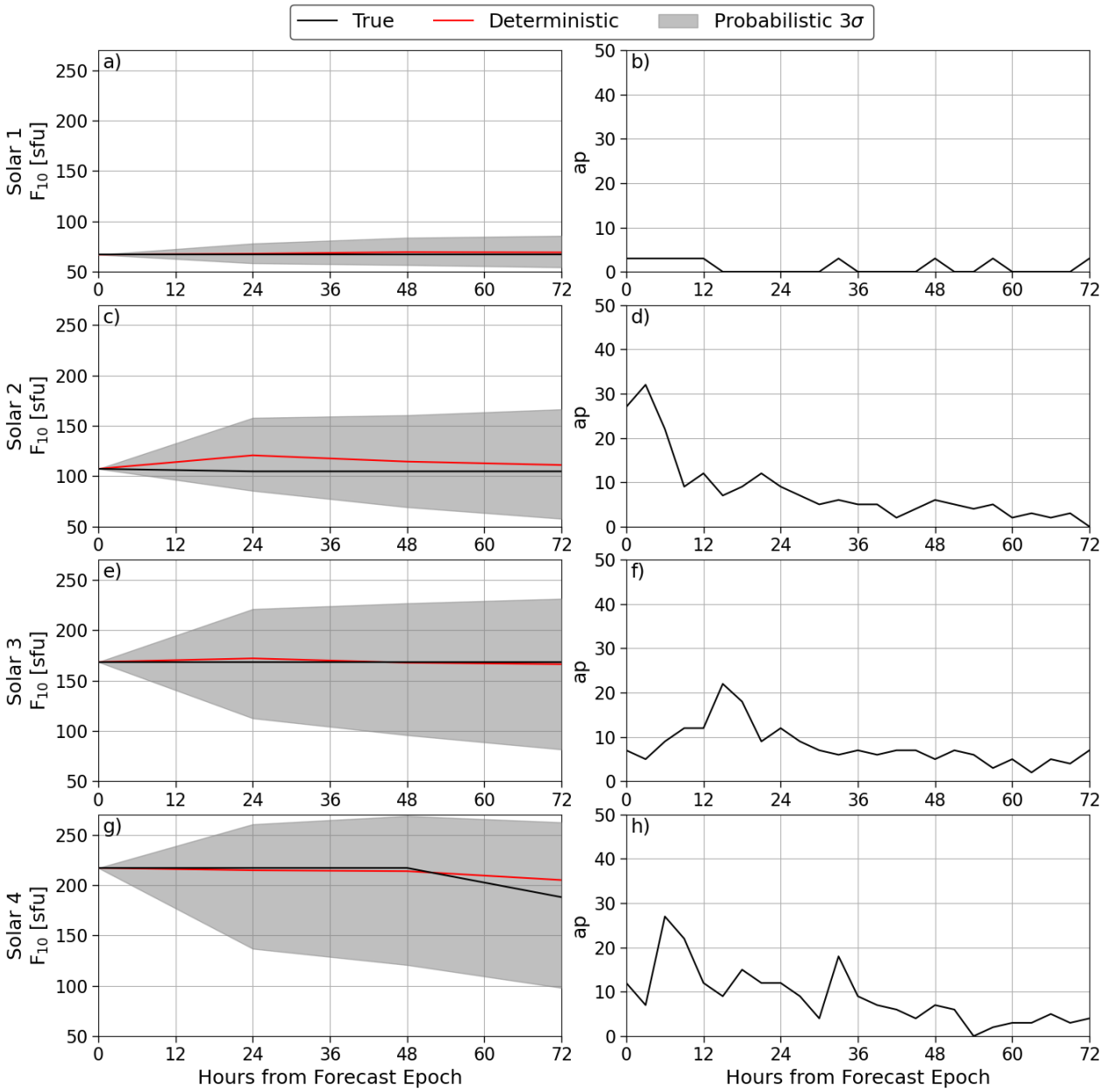

Figure 7.1: Space Weather inputs for the four solar cases. The shaded probabilistic region shows the 3σ bounds for the perturbed samples.

The true, deterministic and perturbed drivers for these periods are displayed in Figure 7.2.

7.1.1 Other Drivers

As previously mentioned, the drivers with available forecasts are the four solar indices, *ap*, and *Dst*. However, TIE-GCM uses *Kp*, MSIS-UQ uses S_N, S_S, ΔT, and *SYM-H*, and CHAMP-ML-v2

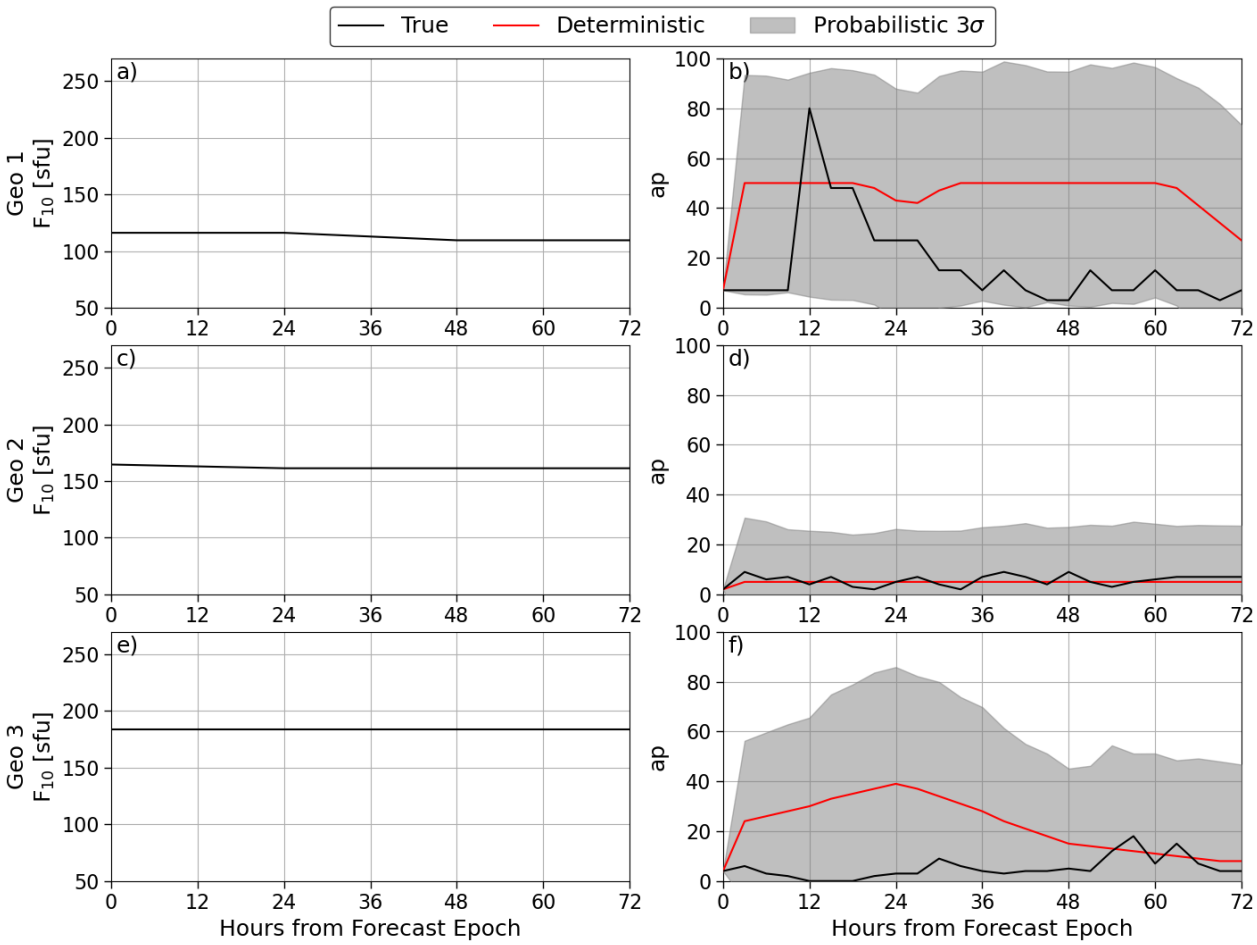

Figure 7.2: Space Weather inputs for the three geomagnetic cases. The shaded probabilistic region shows the 3σ bounds for the perturbed samples.

uses S_N, S_S, and *SYM-H*. To account for this, *Kp* will be transformed from *ap*, and S_N, S_N, ΔT, and *SYM-H* will be derived through a polynomial fit based on *Dst* forecasts. The approximations using *Dst* may not be the ideal approach; however, there are currently no available probabilistic models that forecast these drivers. The polynomial fits for *SYM-H*, S_N, S_S, and ΔT are shown in Figure 7.3.

Linear and cubic fits were also tested for potential transformation equations. However, the quadratic transformations provided the best fit based on a qualitative analysis (e.g. overall fit, fit for extreme values). When using either a true, deterministic, or perturbed *Dst*, estimates for these four geomagnetic drivers can be computed. One additional constraint for S_N, S_S, and ΔT is that they must be positive. All perturbed drivers are kept consistent across the four models.

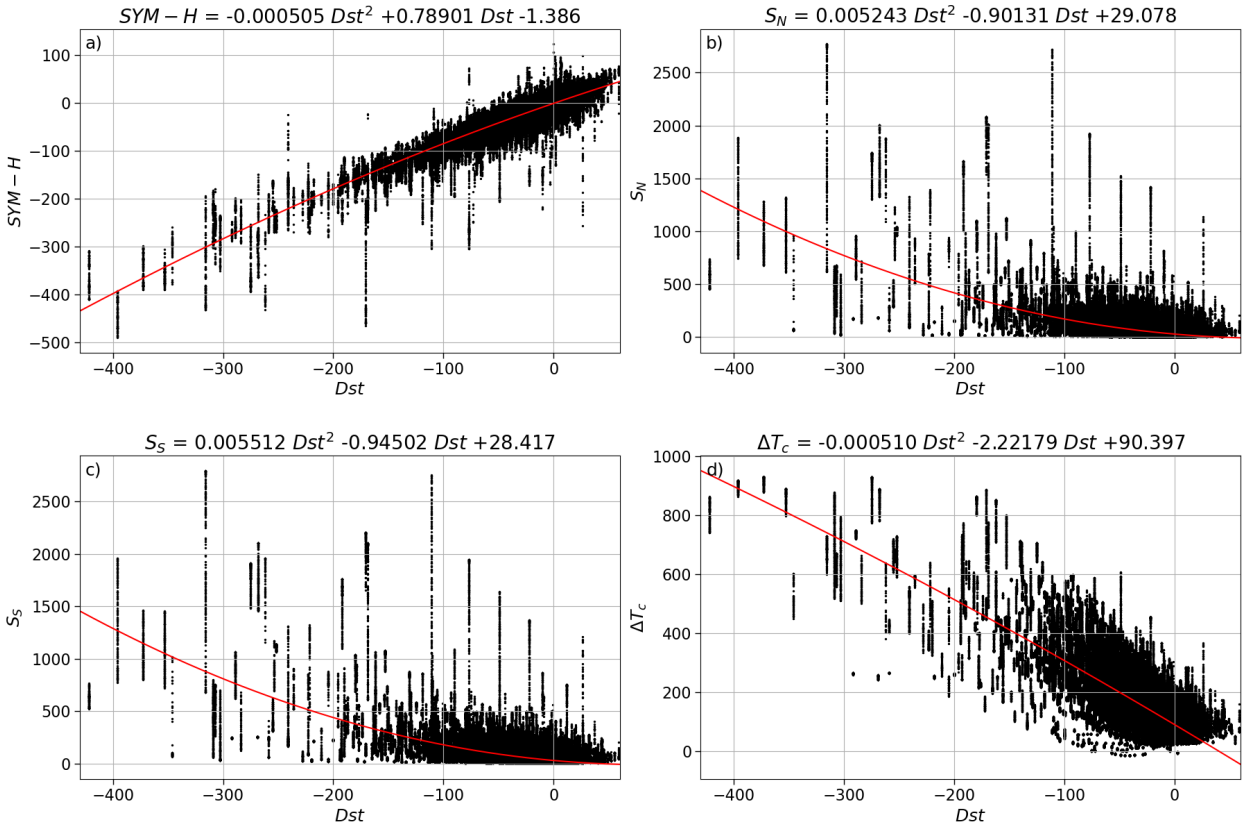

Figure 7.3: Second-order polynomial fits between *Dst* and the four other geomagnetic drivers used by the models.

7.2 Model Uncertainty

To account for model uncertainty, driver variations are deterministic, and probabilistic density is derived from the models' predicted uncertainties. For the solar cases, the deterministic solar drivers are used with the true geomagnetic drivers. For the three geomagnetic cases, the true solar drivers and deterministic geomagnetic drivers are used. This choice was made for consistency with the driver uncertainty cases as the perturbed drivers are based on their respective deterministic forecasts. In these cases, the Monte Carlo runs leverage the models' predictions of μ and σ for thermospheric density. It is important to note that initial state uncertainty (of the thermosphere) is not considered for TIE-GCM ROPE. Since it uses PCA coefficients as inputs, we find periods in the original TIE-GCM dataset with similar starting conditions to each of the seven cases and

use those to initialize the model. In an operational setting, the initial state can be estimated and perturbed through data assimilation.

7.2.1 Density Sampling Approaches

In order to use the density distributions from the models, we must identify the most realistic method to sample density. Here, we will test different approaches starting with traditional MC sampling. This entails directly sampling from the predicted distribution each step of the propagation. However, doing so could potentially negate the effects of density uncertainty, because there is no enforcement of spatiotemporal correlation of density. A quantity used to describe this is the half-life of mass density. Half-life (τ) is the amount of time for the temporal correlation of density to reach 0.50. Commonly-used values for τ in literature are 18 or 180 minutes [140]. A simple approach to combat the uncertainty negation from a direct sampling method is to use a bias factor (κ) to force temporal structure for density. Density would then be defined such that $\rho = \mu + \kappa\sigma$ where μ and σ come from the density model. κ then represents a relative bias in density from the predicted distribution.

This approach requires κ to be sampled from a standard normal distribution every τ minutes. Between these segments, κ would be interpolated with respect to time. This way, the structure of the model distributions are preserved across the Monte Carlo samples without the negation of successive time-steps. The third and final approach is to leverage a first-order Gauss-Markov process [141]. κ is then computed as,

$$\kappa(t+\Delta t) = e^{-\beta\Delta t}\kappa(t) + u_k(t)\sqrt{\frac{\sigma^2}{2\beta}\left(1-e^{-2\beta\Delta t}\right)} \qquad \text{where} \qquad \beta = -\frac{\ln 0.5}{\tau} \tag{7.1}$$

where β is a parameter dependent on half-life, and $u_k(t)$ is a standard normal distribution. The quantity $\frac{\sigma^2}{2\beta}$ represents the steady-state variance of our bias factor, κ, which should be 1.0. The first-order Gauss-Markov process concludes the approaches; however, we will test the second and third approaches with $\tau = 18$ and $\tau = 180$ minutes. We consider a satellite orbit during a generic storm in 2002 using HASDM-ML for the density distributions across the three-day period. Figure

7.4 shows the density for the five cases during the first three hours along with the along-track position difference probability density function (pdf) after the full three-day period.

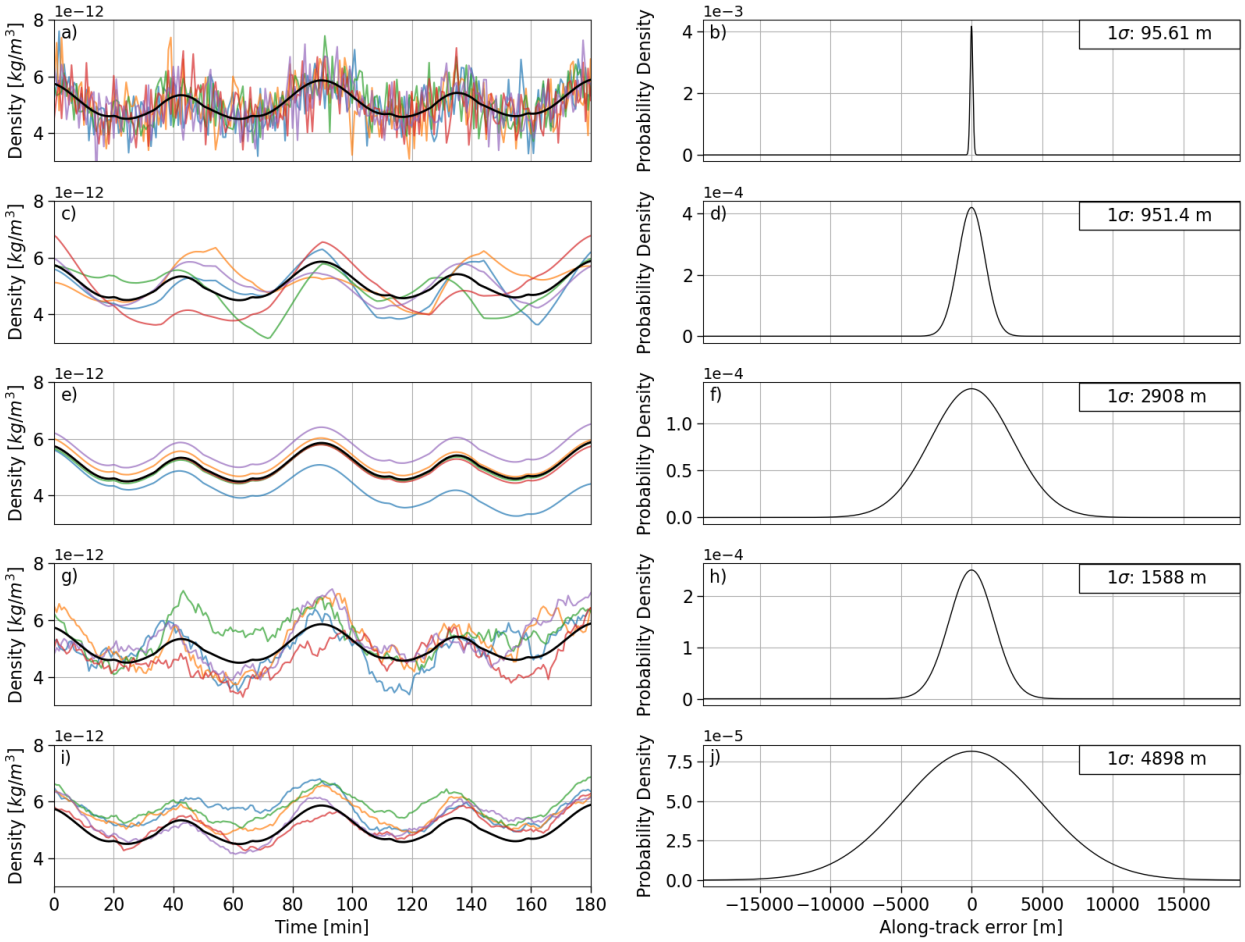

Figure 7.4: Comparison of the Monte Carlo techniques for the traditional Monte Carlo approach (a,b), the interpolated bias factor approach with κ = 18 minutes (c,d) and κ = 180 minutes (e,f), and the first-order Gauss-Markov approach with κ = 18 minutes (g,h) and κ = 180 minutes (i,j). The Monte Carlo mean density is shown alongside density from the first five Monte Carlo runs.

Figure 7.4 panels (a,b) confirm the effect of uncorrelated successive time-steps on position uncertainty. The unstructured density variations cause a severe under-representation of position uncertainty. The second approach – panels (c–f) – shows how forcing a temporal structure in density leads to a large spread of satellite positions. Furthermore, increasing the sampling frequency

($\tau = 18$ to $\tau = 180$) more than triples the standard deviation in the resulting pdf. However, this gradual change in the bias factor does not seem realistic for a given probabilistic run.

The Gauss-Markov process – panels (g–j) – provides a balance between the traditional sampling approach and the basic interpolation for κ. The density variations look more realistic as the temporal structure is preserved while not forcing a given bias. It is difficult to know which half-life is the more-correct choice, as we do not know what the position uncertainty should be for any given period. However, the 18-minute half-life for the first-order Gauss-Markov process seems to provide a more realistic density variation for a Monte Carlo run. The position uncertainty also falls into the middle of the range of values in Figure 7.4. This approach will be used for all density models in the model uncertainty cases for the remainder of this study.

7.3 Results

The following sections are split into the solar conditions (Section 7.3.1) and geomagnetic conditions (Section 7.3.2) to study the effect of the different types of drivers in isolation. It is pertinent to compare the effects of driver and model uncertainty for a given condition.

7.3.1 Solar Activity Conditions

For the first four conditions (Solar 1–4, defined in Table 7.2 and Figure 7.1), we examine the effects of solar activity on position uncertainty. Along-track difference pdfs are shown for all models and both uncertainty sources in Figure 7.5. It is important to note that the reference location for driver uncertainty is the position from HASDM-ML using the mean perturbed drivers, and for model uncertainty, it comes from the HASDM-ML position using only mean density predictions. The pdf statistics can be found in Table 7.3.

To better understand the results in both Figure 7.5 and Table 7.3, the sign convention must be explained. As previously mentioned, HASDM-ML is used to get the reference case for each condition. Every other final position for that condition and uncertainty source is then transformed into the radial, along-track, and cross-track (RTN) frame, and the along-track differences are computed. This refers to how far ahead (positive) or behind (negative) the other positions are with

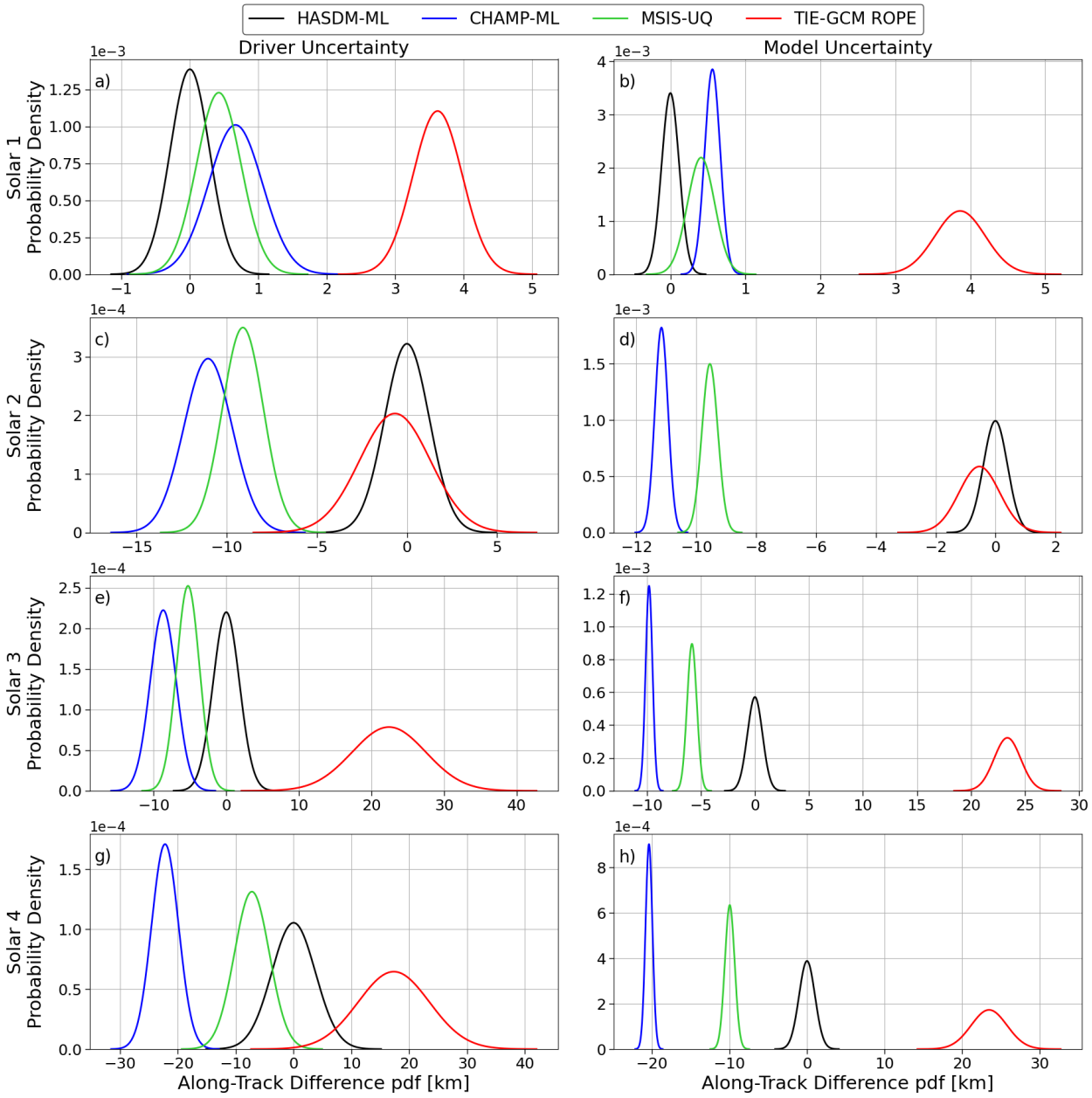

Figure 7.5: In-track position distributions relative to a HASDM-ML reference satellite for the four solar activity conditions in Table 7.2 after 72 hours.

respect to the reference satellite. Biases in the distributions come from persistent biases between model densities for the three-day period based on the conditions. If a satellite encounters higher density relative to another satellite for the same period, it will lose more energy causing a decrease in altitude. This decrease in potential energy leads to an increase in kinetic energy (velocity). The

Table 7.3: Distribution statistics for the four solar activity conditions after 72 hours, corresponding to Figure 7.5.

		HASDM-ML		CHAMP-ML		MSIS-UQ		TIE-GCM ROPE	
Condition	**Uncertainty**	μ (m)	σ (m)	μ (m)	σ (m)	μ (m)	σ (m)	μ (m)	σ (m)
Solar 1	**Driver**	0.0	287.7	666.3	394.9	421.8	324.7	3618	361.2
	Model	-1.5	117.2	559.3	103.6	408.2	182.0	3864	335.6
Solar 2	**Driver**	0.0	1239	-11043	1345	-9108	1141	-665.1	1965
	Model	-6.2	401.7	-11165	218.9	-9546	266.0	-541.6	679.3
Solar 3	**Driver**	0.1	1815	-8690	1795	-5272	1582	22419	5083
	Model	-9.2	689.2	-9820	319.6	-5846	445.6	23377	1235
Solar 4	**Driver**	0.3	3784	-22226	2333	-7210	3036	17326	6172
	Model	13.7	1028	-20368	442.1	-9948	629.0	23492	2310

satellite experiencing higher density – therefore more drag – will end up ahead with a positive along-track position difference.

In Figure 7.5 panel (a), the solar minimum conditions result in a total position spread of approximately 6 km across the four models. The distributions for HASDM-ML, CHAMP-ML and MSIS-UQ have significant overlap indicating they are predicting similar density levels with the perturbed solar drivers. All models in this case have similar sensitivity to the uncertainty in solar drivers as the distributions have similar spreads. The standard deviations are logged in Table 7.3. The only model with a notable bias for this low solar activity condition is TIE-GCM ROPE. This bias indicates a persistent over-prediction of density relative to the other models. Shifting to panel (b), model uncertainty is visibly less impactful on the position distributions. The biases are similar to the driver uncertainty case, but the standard deviation of each distribution is 2–3 times smaller when considering only model uncertainty. This is most pronounced for CHAMP-ML with σ being 349.9 m and 103.6 m for driver and model uncertainty, respectively. The exception to this is TIE-GCM ROPE where the σ is similar after 72 hours.

The Solar 2 case (moderate solar activity) results in significantly larger position spreads, seen in panels (c,d). For this condition, TIE-GCM ROPE now has a small bias relative to HASDM-ML while CHAMP-ML and MSIS-UQ considerably under-predict density (in a relative sense). The

distributions are now much wider in panel (c) than they were in panel (a) which could be explained by the increased spread in the solar drivers for this case (refer to Figure 7.1). The driver uncertainty for TIE-GCM ROPE for moderate solar activity is now noticeably larger than for the other models. Comparing driver to model uncertainty for this condition, the disparity is much more pronounced. The uncertainty for perturbed solar drivers now causes 3–6 times larger uncertainty than the use of probabilistic models. Another key result for Solar 2 is the difference in model uncertainty for the two localized models (CHAMP-ML and MSIS-UQ) compared to the two PCA-based ROMs (HASDM-ML and TIE-GCM ROPE).

For the elevated solar activity case (panels (e,f)), the overall position spread increases to nearly 60 km and 35 km for driver and model uncertainty, respectively. For both uncertainty sources, TIE-GCM ROPE produces the largest uncertainties of the four models with a 5 km standard deviation in the along-track direction. TIE-GCM ROPE is again predicting much higher density than the other three models. The final condition, high solar activity (panels (g,h)), shows fairly similar results. The ordering of the four models (in terms of the bias) is the same as the elevated solar activity case. TIE-GCM ROPE again has the largest position spread for both uncertainty sources, but the disparity is not as great as it was for elevated solar activity. The comparison between driver and model uncertainty is also very similar to the previous condition for each of the models. The notable result for high solar activity, though, is that the total position spread is 70 km which highlights an issue raised in Chapter 1. The choice of model when performing PoC calculations has a strong impact on the results.

Across the four solar activity conditions, driver uncertainty was typically 2–6 times larger than model uncertainty for a given thermosphere model. While it can seem significant in some cases, model uncertainty is certainly not negligible. This is made more apparent in Figure 7.6 which shows the along-track difference standard deviation as a function of time for all models and conditions examined in this section.

Figure 7.6 provides us with insight into the relative importance of the uncertainty sources in the first 36 hours of the satellite propagations for the different solar activity conditions. Note that the y-

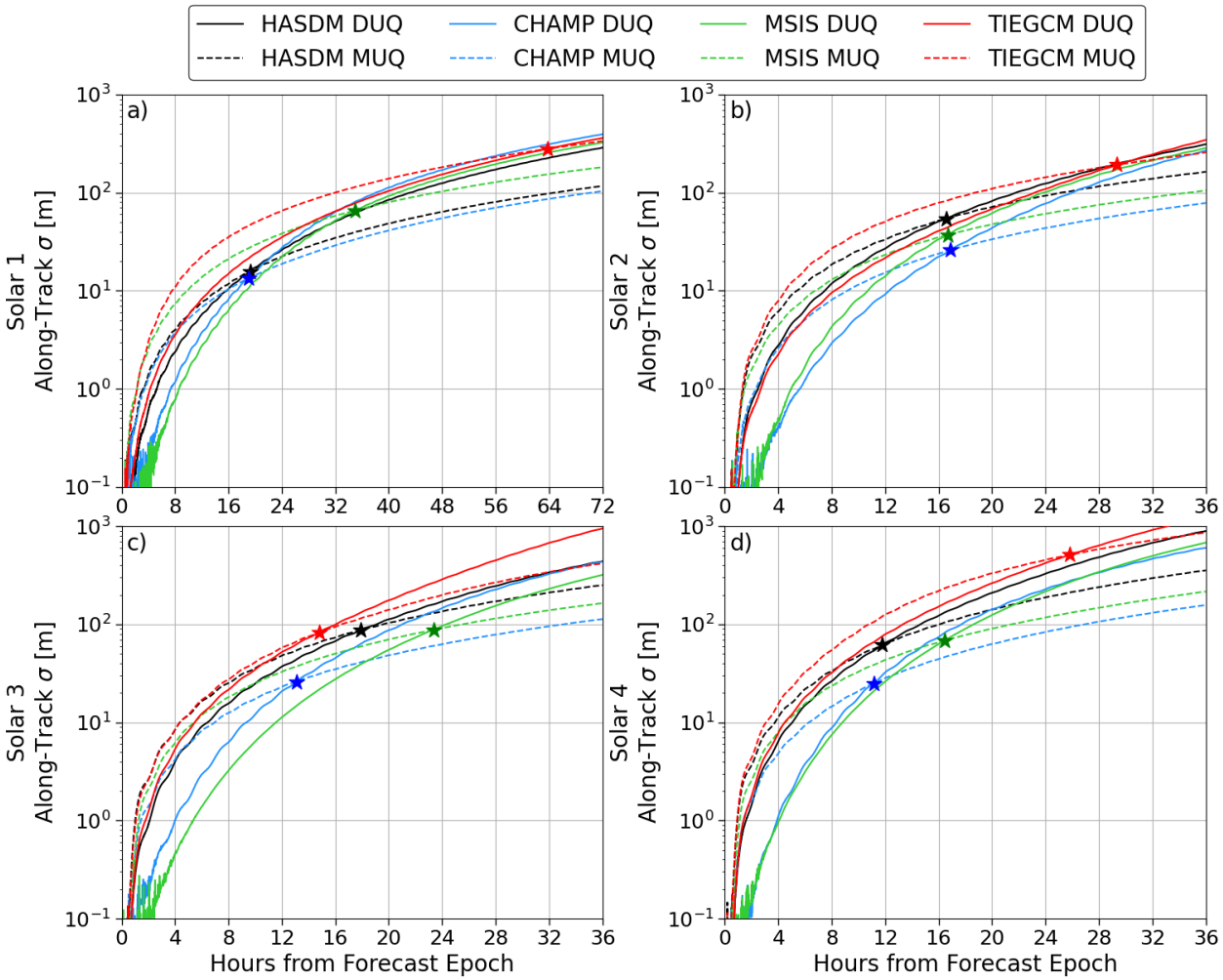

Figure 7.6: In-track position standard deviation as a function for time for the four solar activity conditions in Table 7.2. The markers refer to the time where the dominant uncertainty takes over for a particular model.

axes are in log-scale to highlight differences early in the periods where the standard deviations can be both similar and small in magnitude. For solar minimum (panel (a)), model uncertainty results in the larger position spread for all models in the first 18 hours. At this point, driver uncertainty takes over as the dominant source for both HASDM-ML and CHAMP-ML. It takes over 34 hours for this convergence point using MSIS-UQ. For TIE-GCM ROPE, this does not occur until after approximately 64 hours. In panel (b), we see an interesting occurrence at around 16 hours. It is at this point that the two uncertainty sources converge for HASDM-ML, CHAMP-ML, and MSIS-UQ. This transition occurs for TIE-GCM ROPE right before the 30-hour mark.

For elevated solar activity (panel (c)), the transition time is more spread between the models. The transition occurs early for both TIE-GCM ROPE and CHAMP-ML, both within the first 15 hours. HASDM-ML remains fairly consistent at about 17 hours, and MSIS-UQ take approximately one day for driver uncertainty to take over. Panel (d) shows a quick rise in uncertainties and driver uncertainty takes over for HASDM-ML, CHAMP-ML, and MSIS-UQ within the first 17 hours. It takes around 26 hours for the transition to occur for TIE-GCM ROPE.

These results highlight the importance of both uncertainty sources in conjunction assessment. After 72 hours, the differences are quite considerable, but in the first 12–36 hours, both uncertainty sources have similar impacts on orbit prediction. In an operational setting, when there is a potential conjunction forecasted, the assessment is repeatedly updated as it nears to see if and how the outlook has changed. If model uncertainty is ignored, there is important information being left out as the potential conjunction approaches.

7.3.2 Geomagnetic Activity Conditions

For the last three conditions (Geo 1–3, defined in Table 7.2 and Figure 7.2), we examine the effects of geomagnetic activity on position uncertainty. Along-track difference pdfs are again shown for all models and both uncertainty sources in Figure 7.7. The pdf statistics corresponding to Figure 7.7 can be found in Table 7.4.

Table 7.4: Distribution statistics for the three geomagnetic activity conditions after 72 hours, corresponding to Figure 7.7.

		HASDM-ML		**CHAMP-ML**		**MSIS-UQ**		**TIE-GCM ROPE**	
Condition	**Uncertainty**	μ **(m)**	σ **(m)**	μ **(m)**	σ **(m)**	μ **(m)**	σ **(m)**	μ **(m)**	σ **(m)**
Geo 1	**Driver**	0.0	1186	-9846	963.5	-14557	672.5	1443	1692
	Model	-15.0	441.5	-11064	603.9	-15063	382.9	3607	1553
Geo 2	**Driver**	0.1	2375	-11235	1599	-4683	946.0	-16973	2528
	Model	5.1	1043	-14254	637.9	-5134	748.1	-16354	828.0
Geo 3	**Driver**	0.1	2210	-13029	3478	-8798	1278	19292	3366
	Model	8.7	1250	-21852	923.1	-13011	809.7	28662	5098

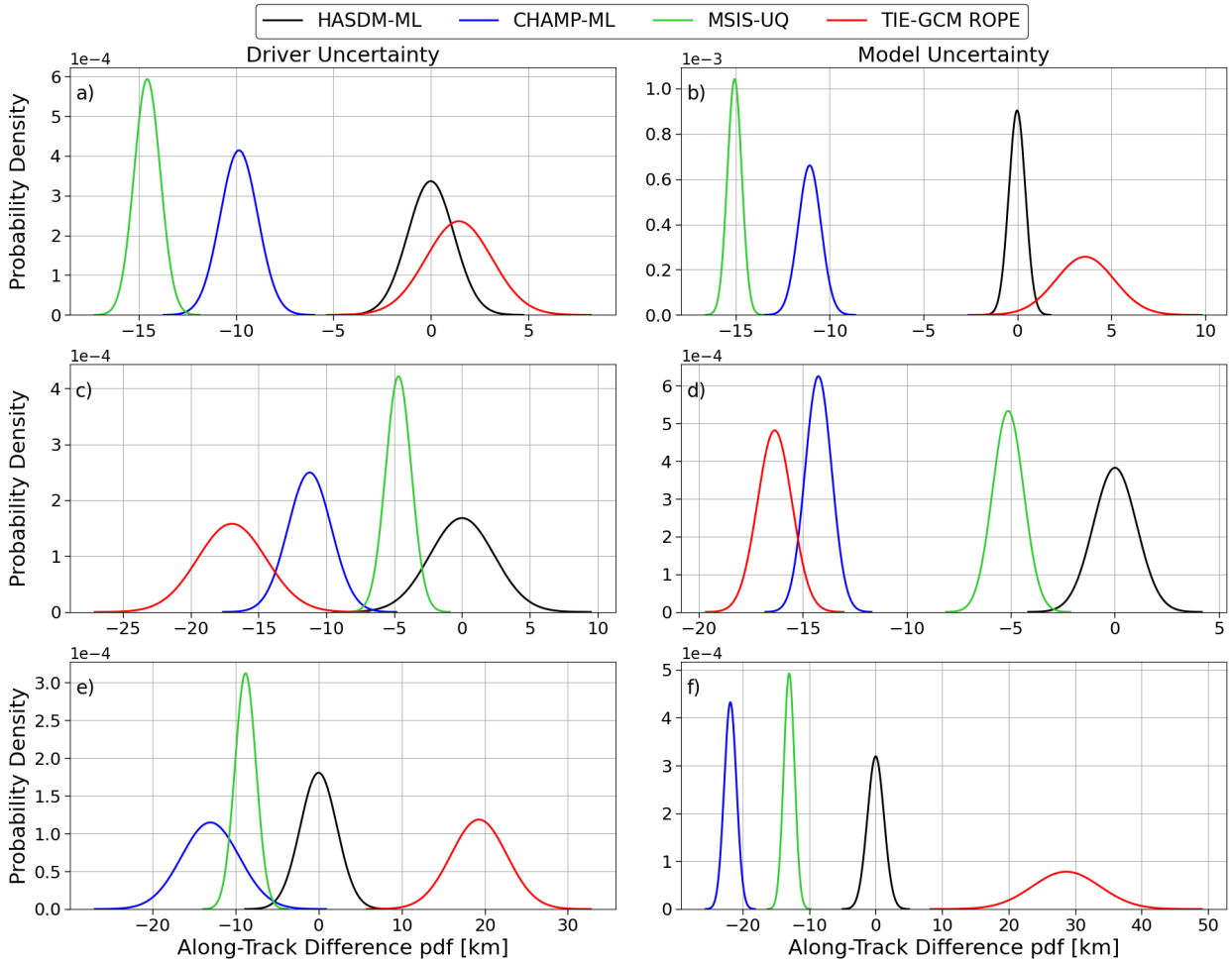

Figure 7.7: In-track position distributions relative to a HASDM-ML reference satellite for the three geomagnetic activity conditions in Table 7.2 after 72 hours.

The Geo 1 condition in Figure 7.7 panels (a,b) represents moderate solar and geomagnetic activity. The forecast for *ap* (Figure 7.2) shows that the deterministic forecasts remains mostly at the threshold between moderate and active geomagnetic activity, and the 3σ range for *ap* is between 0 and 100 . This results in along-track σ between 670 and 1700 m across the models for driver uncertainty. This is around the levels for the Solar 2 driver uncertainty case where the solar indices were uncertain. However, that case had lower geomagnetic activity indicating that the models have higher sensitivity to solar drivers. In this condition, TIE-GCM ROPE shows to have the largest driver and model uncertainty while also having similar density levels to HASDM-ML (small bias).

The results in panel (c) are particularly interesting given the low uncertainty in geomagnetic activity for this period. Referring back to Figure 7.2, Geo 2 has consistently low magnitude *ap* with 3σ uncertainty bounds hardly reaching 30 while Geo 1 was considerably more active and uncertain. The true F_{10} for Geo 1 stays around 120 sfu while it remains fairly constant around 160 sfu for Geo 2, although we do not consider any solar driver uncertainty in these cases. Comparing panels (a) and (c) from Figure 7.7, there is much more driver uncertainty for Geo 2. The along-track σ is 1.5–2 times larger, and the overall position spread increases from 25 km to over 35 km. This highlights how important solar activity level is to these density models.

For the final case, seen in panels (e,f), considers high solar activity and moderate geomagnetic activity. Here, we see an overall position spread of approximately 60 km and 70 km for driver and model uncertainty, respectively. The along-track σ for driver uncertainty ranges from 1.2–3.5 km while ranging from 0.8–5.1 km for model uncertainty. Again, the general trend is that driver uncertainty is more impactful than model uncertainty but with its first exception. For TIE-GCM ROPE, model uncertainty (σ) is actually larger by over 1.7 km. As this case considers moderate geomagnetic activity, we can compare these results to Geo 1. In that case, the TIE-GCM ROPE driver uncertainty σ was only slightly larger than for model uncertainty. Granted, the geomagnetic drivers were more uncertain in that period, the discrepancy could be caused by the model's response to higher geomagnetic activity during higher solar activity.

Once more, we investigate the growth of along-track σ for each model from both uncertainty sources (similar to Figure 7.6). These results are shown in Figure 7.8 for the three geomagnetic cases. Note: Geo 1 is shown twice in panels (a) and (d), where panel (d) shows the full 72-hour period.

In Figure 7.8 panel (a), the transition from model to driver uncertainty takes place early (before the 14-hour mark) for HASDM-ML, CHAMP-ML, and MSIS-UQ. When comparing these results to Figure 7.6, it is important to consider the different x-axis limits. Looking at panel (d), we see that the transition occurs for TIE-GCM ROPE after 56 hours which is similar to Figure 7.6 panel (a). Figure 7.8 Panel (b) again shows that by 13 hours into the propagation, driver uncertainty

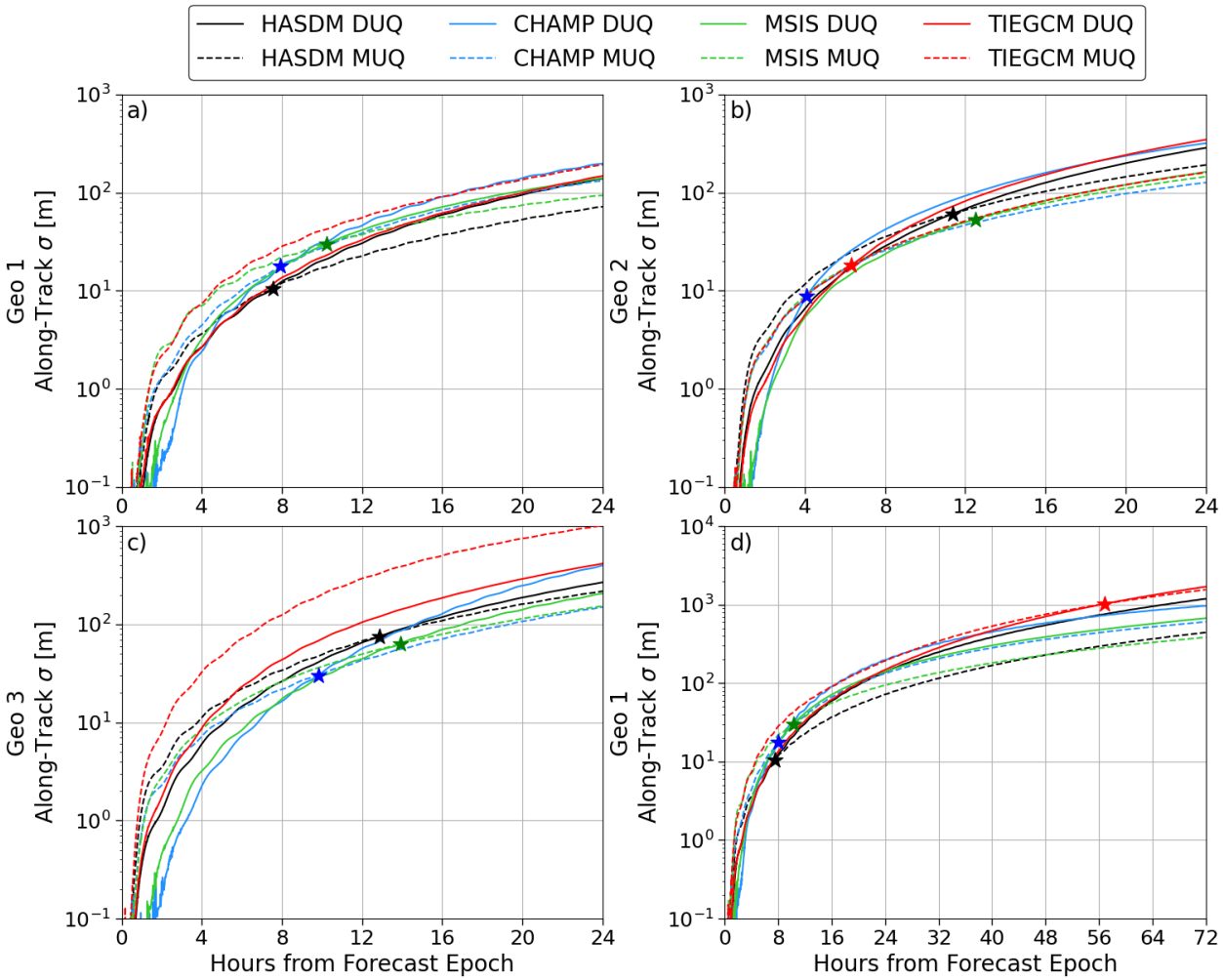

Figure 7.8: In-track position standard deviation as a function for time for the three geomagnetic activity conditions in Table 7.2. The markers refer to the time where the dominant uncertainty takes over for a particular model. Note: panels (a,d) are both for the "Geo 1" condition, but (d) provides in-track σ for the full 72-hour propagation period.

is the dominant source. Here, driver uncertainty takes over after only six hours for TIE-GCM ROPE and CHAMP-ML. The results for Geo 3 (panels (c,d)) do diverge from the previous cases. For HASDM-ML, CHAMP-ML, and MSIS-UQ, the transition occurs between 9 and 14 hours. However, as seen prominently in Figure 7.7, the transition does not occur at all for TIE-GCM ROPE in this period.

7.4 Summary

This chapter is the culmination of all the work in this dissertation. The four thermosphere models developed are applied to a very important problem related to STM: determining the impact of driver and model uncertainty on a satellite's state. Current operations either do not account for these uncertainties or over-simplify them. We investigate driver uncertainty by perturbing deterministic forecasts with error statistics generated in Chapter 6. This results in along-track standard deviations as high as 6.1 km.

For model uncertainty, deterministic drivers are used with the probabilistic thermosphere models. A first-order Gauss-Markov process allows us to realistically sample from the model distributions for density. The along-track standard deviations reach upwards of 5 km. While model uncertainty does not seem to affect the satellite state as severely, we show that in the first 4–36 hours, it is more impactful that driver uncertainty. In one case, we observed model uncertainty having a larger impact across the entire 72-hour period.

Chapter 8. Summary and Conclusions

In the modern space age, rockets launch on a weekly basis carrying dozens of satellites to reside in low Earth orbit. This is now driven heavily by the commercial industry. We can no longer rely on simply cataloguing orbiting objects and tracking their states. The proliferation of LEO has led to a recent and crucial focus on space traffic management. The current needs are accurately predicting the orbits of all objects and focusing on potential satellite collisions. Short-term forecasting – on the order of hours or days – is critical to prevent potential collisions, which are catastrophic to the space environment. The 2009 Iridium-Cosmos collision resulted in approximately 2,300 observable debris objects, 65% of which remained in orbit seven years later [142]. Debris objects created by collisions or weapons tests can catapult into highly elliptical orbits, which pose a danger to satellites in multiple orbital regimes [143].

Currently, we do not have the capability to either adequately predict model drivers for thermospheric mass density or to predict density perfectly given a set of space weather conditions. Therefore, satellite drag carries uncertainty that varies as a function of location, time, and space weather conditions. Currently, this uncertainty is ignored, or it is reduced to some basic representation. This can lead to a lack of confidence in decision-making when a potential collision is detected. Operators need to have an accurate representation of future satellite states in order to determine if an avoidance maneuver is imperative, as these decisions are expensive.

This dissertation focused on the novel development of probabilistic thermosphere models and uncertainties in a satellite's state. **The development of HASDM-ML, CHAMP-ML, MSIS-UQ, and TIE-GCM ROPE marks the first time probabilistic thermosphere models were developed with demonstrated reliability of uncertainty estimation capabilities and is a major contribution of this work.**

HASDM-ML is a reduced order surrogate model for the SET HASDM density database with robust and reliable uncertainty estimation capabilities. Principal component analysis 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 using Monte Carlo dropout. 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 early version of 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. We 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.

Upon selecting HASDM-ML, we evaluated its uncertainty capabilities across the entire orbits of CHAMP and GRACE-A, 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 dataset. 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. The results from the density calibration tests are significant, because probabilistic thermospheric density modeling is still a novel concept. Additionally, uncertainty estimates themselves are not 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.

HASDM-ML was further improved by the incorporation of a direct probability distribution prediction approach. This also made the model much more computationally efficient. This iteration of the model had better performance across all space weather conditions tested (through the binning of F_{10} and ap). Additionally, we examined spatial calibration maps across the test set to show how the calibration varied laterally and with altitude. This provided insight into the model's tendency to

underestimate uncertainty at the lowest altitudes (where there is not as much variability) although it was more calibrated from 400–800 km.

CHAMP-ML was developed on the in-situ CHAMP accelerometer-derived density dataset from Mehta et al. [48]. The model was first developed using both MC dropout and direct probability prediction for comparison with the HASDM-ML results. The two approaches yielded similar models, although the direct probability method was more accurate. Additionally, computational constraints made MC dropout difficult for the 25 million sample dataset. We tested CHAMP-ML's global prediction capabilities by generating baseline maps during the winter and summer solstices to ensure global physical trends are being captured by the CHAMP model. This showed that the model was able to emulate the effect of Earth's tilt.

We also performed global evaluations for eight unique conditions to determine the altitude dependence of model uncertainty. The altitude profiles showed that the minimum and maximum 1-σ uncertainties were 10 – 28% of the mean predictions, respectively. Solar activity was most influential in determining the profiles' shapes, while geomagnetic activity and the day of year tended to provide uniform changes in the uncertainty. These uncertainty profiles confirmed that the uncertainty estimates were indicative of the original dataset, where uncertainty was elevated in condition-location combinations that were absent or minimal in the dataset. CHAMP-ML was later improved with the introduction of time-series *SYM-H* inputs. Although this decision was made with storm and post-storm conditions in mind, it reduced the overall error by approximately 1.5%.

MSIS-UQ is another probabilistic model developed on in-situ measurements, but it predicts exospheric temperature as opposed to mass density. The exospheric temperature dataset it was developed on was generated such that the temperatures could be input to NRLMSIS 2.0 to match density estimates from the CHAMP, GRACE, and Swarm satellites. Therefore, this was a quasi-correction model, but it crucially transformed NRLMSIS 2.0 into a probabilistic thermosphere model. We showed that across all 81 million samples in the original dataset, MSIS-UQ reduced the error from NRLMSIS 2.0 and HASDM by 28% and 12%, respectively. It was also well-calibrated

across the training, validation, and test sets.

The uncertainty estimates were closely examined for a given time where CHAMP and GRACE were at unique locations in terms of local time and latitude. The uncertainty bounds for species densities showed potential for scientific value when considering relative abundances or the uncertainty associated with the O-He transition region. Instead of having a specific altitude where He takes over as the dominant constituent, we observed a 1-σ interval of 45-50 km where this may occur, depending on geographical location. This study also highlights the improvement in temperature and density prediction with the MSIS-UQ T_∞ predictions. Not only is the bias reduced, the uncertainty estimates can be used to inform decision-making. We observed the effect of uncertain exospheric temperature on the relative uncertainty in density as a function of altitude, highlighting the ability to provide different uncertainty ranges as a function of position.

The final model developed as part of this work was TIE-GCM ROPE. This is an ensemble of ten LSTMs trained on seven years of TIE-GCM outputs during solar cycle 23. The direct probability method for UQ used in the other models proved to be ineffective for this dynamic modeling application, which spurred the ensemble approach. A fixed weighting scheme was determined based on individual model performance throughout the training set to get a weighted mean prediction based on observed errors. We use the ensemble predictions to extract a sample standard deviation and use a scaling method to ensure better calibration of the uncertainty estimates.

TIE-GCM ROPE was compared to Dynamic Mode Decomposition with control, a popular dynamic ROM technique for the thermosphere. We show that while general errors are virtually equivalent for the two approaches, TIE-GCM ROPE does considerably better in capturing the dynamics of TIE-GCM. It also follows the data for extreme periods and has the capability to emulate the system, making dynamic predictions (no state update) for up to a year or over 8,700 time steps with approximately 10% mean error. The uncertainty estimates are not *as* reliable as the other probabilistic models developed here, but it is an important capability due to the difficulty in UQ for physics-based models.

The second contribution is a novel study to extract science through machine learning with

a focus post-storm thermospheric overcooling. Using NRLMSIS 2.0, HASDM-ML, CHAMP-ML, MSIS-UQ, and JB2008-ML (developed solely for this study), the inherent relationship between geomagnetic activity and density is quantified within the respective datasets or models they are based on. The study is motivated by considering the infamous 2003 Halloween storms. This storm was a result of successive coronal mass ejections that substantially increased thermospheric density twice in two days. This event disturbed the thermosphere for multiple days, and when the storm was over, density was over 25% below its pre-storm levels. This event is a true test for thermosphere models during extreme events.

By looking at orbit-average density over this period, we were able to see that both NRLMSIS 2.0 and JB2008-ML could not model the sudden drop in density while the other three models better matched the data. However, there are other factors that impact the thermosphere in reality, and it is difficult to isolate the effect of geomagnetic activity. To overcome this challenge, the models' time series geomagnetic drivers were independently varied while holding all other drivers constant. This showed that NRLMSIS 2.0 has a fairly linear relationship with geomagnetic activity, and no drivers resulted in a sub-1.0 density ratio (which would indicate overcooling). JB2008-ML also did not exhibit this phenomenon. In fact, the most important historical driver of JB2008-ML was the 9-hour prior *ap* which resulted in density ratios nearly twice that of any other driver. This is due to a 9-hour offset in the calculation of the JB2008 parameter DTC based upon *ap* use (Bowman and Tobiska, private communication, 2020).

HASDM-ML was most strongly driven by the current and 6-hour prior *ap* for thermospheric expansion, while increases in $ap_{12\text{-}33}$ and $ap_{36\text{-}57}$ resulted in densities as low as 57% of the baseline magnitude. CHAMP-ML indicated a highly nonlinear relationship between density and geomagnetic activity. Depending on the location, *SYM-H* or $SYM\text{-}H_{0\text{-}3}$ drove the largest density ratios, significantly more than any other model. In terms of cooling, CHAMP-ML showed that at *SYM-H* > -100 nT, many of the recent drivers caused a density ratio of less than 1.00. As the index was made more negative, the least recent drivers caused the lowest density ratios, particularly at low latitudes. MSIS-UQ also showed a nonlinear relationship between geomagnetic activity and den-

sity, although its density ratios were not as extreme. At high-latitude, MSIS-UQ had density ratios below 0.75 for the least recent geomagnetic drivers while ap_3 caused the largest positive ratios at the equator.

While thermosphere model performance is important, forecasting density relies heavily on our ability to forecast the model drivers. **The third contribution of this dissertation is benchmarking the current operational forecasting models that drive JB2008 and the HASDM system.** The analysis of the SET algorithms used by the JB2008 and HASDM models provided clear performance baselines for the current state-of-the-art of operational, automated density model driver forecasts. This work showed the strengths of these predictive algorithms while also showing conditions where improvements can be made. In general, the forecasting capability for solar indices at low and moderate activity levels has comparably low uncertainty and virtually no bias. This performance is degraded to an extent at elevated and especially high activity levels, where the Sun is more volatile, and the evolution of flaring active regions is still poorly predicted.

The best performing solar driver algorithm is for Y_{10}, whose forecasting method is the most complex of the four solar indices investigated. The algorithm for M_{10} also has low uncertainty and low bias at the two lower solar activity levels. The forecasts for F_{10} and S_{10} prove to be more uncertain and with generally higher biases. Both indices had strong tendencies to over-predict at high solar activity. The index that delivers the greatest energy input to the atmosphere is S_{10}, so reducing the error in this driver would significantly improve density forecasting overall.

The geomagnetic indices, *ap* and *Dst*, proved to be difficult to predict even using two diverse methods. The forecasts for *ap* are determined by a team of forecasters with the aid of a model, and there was still a low probability of detection for geomagnetic storms. In most conditions however, there was little or no bias in the predictions. The three-day prediction window also ended up being a limitation, and results from a full six-day forecast would be intriguing. The *Dst* algorithm performed well during G0 (or quiet) conditions. The standard deviation of error stayed steady at around 13 nT in these cases. The algorithm showed poorer trends with increased geomagnetic activity. The increased uncertainty is attributed to the lack of HSS prediction and an inability to

accurately and consistently forecast CME arrival time and magnitude.

Combining the preceding work, the fourth and final contribution is the analysis of the impact of driver and model uncertainty on a satellite's state. Using the error statistics from Chapter 6, we can perturb deterministic space weather driver forecasts. These perturbed drivers are used in Monte Carlo analyses to determine uncertainty bounds for a satellite's state. Additional model drivers used in the probabilistic thermosphere models (not covered in Chapter 6) do not have operational forecasts. To overcome this, we fit polynomials to transform *Dst* forecasts to all other geomagnetic drivers required to use these models. The probabilistic models can provide density distributions at any point in a satellite orbit, paving way for model uncertainty to be incorporated into satellite orbit propagation. Using a first-order Gauss-Markov process with an 18-minute half-life, we are able to generate density samples from the predicted distribution while maintaining temporal correlation.

The goal of this study is to consider driver and model uncertainty independently and see how these prominent sources affect a satellite's state along a 72-hour orbit. Seven time periods are chosen based on space weather conditions. The first four cover the different levels of solar activity. For these, the driver uncertainty cases will only consider perturbed solar driver forecasts in addition to model uncertainty. The last three periods cover different combinations of solar and geomagnetic activity where only the geomagnetic driver forecasts are perturbed.

For the first four periods, we observe a clear relationship between solar activity and satellite position distributions across all models. As solar activity increased, the position spreads increased within and across the models. For driver uncertainty, the along-track σ was larger, and many of the model distributions overlap. The relative biases change across the different periods. For model uncertainty, the spread using each model was notably smaller, and the distributions were often disjoint. Looking at the growth of the along-track σ for each model and uncertainty source showed that although driver uncertainty is dominant after 72 hours, model uncertainty is the larger contributor to uncertainty in the orbital state for the first 8–36 hours.

The last three periods displayed many of the same observations for model uncertainty, although

geomagnetic driver uncertainty resulted in different state uncertainties. There is no clear determination of whether solar or geomagnetic driver uncertainty has more of an impact on the orbital state since the along-track σ for driver uncertainty was inconsistent between the same models for similar conditions (e.g. Solar 2 and Geo 1, Solar 3 and Geo 2). As with the first four periods, model uncertainty was again more important early on in the propagation but eventually becomes second to driver uncertainty. In these cases, the transition generally occurred earlier (between 4 and 14 hours) although for TIE-GCM ROPE during the final period, model uncertainty remained dominant throughout the 72 hours.

This study highlighted the importance of considering both driver **and** model uncertainty in conjunction assessment for operations. As a potential conjunction event nears, model uncertainty will become dominant, and it is currently not considered in operations. Even though driver uncertainty became dominant within the first day for most models and periods, model uncertainty did not become negligible. Probabilistic thermosphere models must be a focus not only in future modeling efforts, but also in the operational framework for the future of space traffic management.

8.1 Future Work and Recommendations

This work has brought light to the importance of probabilistic modeling for thermospheric density and the proper treatment of uncertainty in conjunction assessment. However, considerable work is still required to address the challenges of the modern space age. The next steps and recommendations for the continuation of this work are as follows.

Nonlinear Dimensionality Reduction: HASDM-ML and TIE-GCM ROPE are reduced order models meaning they operate in a reduced state. In this work, PCA provided the transformations between the full and reduced states which made UQ achievable. PCA is an optimal linear technique, but we know that the thermosphere can become highly nonlinear during geomagnetic storms. This causes truncation errors to rise during these important events. Nonlinear approaches (e.g. kernel PCA, convolutional autoencoders, bidirectional generative adversarial networks) may provide an avenue for improved mapping between the the thermosphere and the reduced state that the models operate in.

Expand Data-Driven Modeling Efforts: CHAMP-ML was the one model developed in this work entirely based on observations. This bypassed all predefined assumptions about the system and was able to predict meaningful density distributions on a global scale. The major drawback to this model is the spatiotemporal window it covered. This modeling approach combined with additional satellite data would improve the applicable altitude range and offer multiple observations on individual events. Other satellite missions with similar data include: GRACE A and B, GRACE-FO, Swarm A, B, and C, and GOCE.

Additional Outputs for MSIS-UQ: NRLMSIS 2.0 formulates a temperature profile using three parameters: the temperature at 120 km (T_{120}), the vertical gradient at 120 km (T'_{120}), and the exospheric temperature (T_∞). MSIS-UQ focused on the T_∞ aspect of temperature profiles which showed to be effective in calibrating NRLMSIS 2.0 to satellite data and make it probabilistic. However, the exospheric temperature has little impact on density below 150 km. The numerical approach for obtaining the corrected T_∞ estimates can be extended to get estimates for T_{120} and T'_{120}. This would be particularly useful in incorporating uncertainty to re-entry applications. A major challenge with this task, however, is simultaneously estimating the three parameters.

Development of Probabilistic Driver Forecast Algorithms/Models: The error statistics obtained from the benchmarking study proved to be useful for perturbing deterministic driver forecasts to enable the consideration of driver uncertainty. Nevertheless, we showed how important probabilistic modeling was for the thermospheric density application. Future development of driver forecast algorithms should adopt this approach and incorporate uncertainty techniques. This would be crucial for improving current operational methods with regards to space weather driver uncertainty.

Probabilistic Driver–Density Coupling: Building off of probabilistic driver forecast algorithms and models, the MC approach for driver uncertainty in orbit prediction is generally inefficient. There should be a future focus into model development that can transform distributions of space weather drivers into probabilities in density. This would likely require either (1) advanced modeling techniques or (2) carefully procured data for a model to be able to take probabilistic inputs and reliably map it to thermospheric density. If accomplished, the density model would be in-

valuable as it would lump the two major uncertainties into a single efficient package for improved operations.

Probabilistic Multi-model Density Ensemble: With the development of these probabilistic thermosphere models, we can now consider the state of an object in orbit as a combination of the states from different models. Each model has strengths, and there is currently no way of knowing which is going to be the most accurate for a given period. The conjunction assessment may become more complicated, but it provides a better picture of the uncertainty of an object given the different possible evolutions of the thermosphere.

Additional Thermospheric Density Data: Over the past two decades, we have been receiving new datasets from which we can extract thermospheric mass density. This data is invaluable for thermosphere model development and evaluation. The continued launch of satellite missions with high-fidelity accelerometers and GPS are crucial for the continued improvement of both our understanding of the thermosphere and the models themselves. This needs to remain a major focus. The commercial megaconstellations could provide a good avenue to get more global measurements of the thermosphere.

References

[1] United States Space Command. *Satellite Catalogue*. 2022. URL: https://www.space-track.org.

[2] Theodore J. Muelhaupt et al. "Space traffic management in the new space era". In: *Journal of Space Safety Engineering* 6.2 (2019), pp. 80–87. ISSN: 2468-8967. DOI: https://doi.org/10.1016/j.jsse.2019.05.007. URL: https://www.sciencedirect.com/science/article/pii/S246889671930045X.

[3] Hanspeter Schaub et al. "Cost and risk assessment for spacecraft operation decisions caused by the space debris environment". In: *Acta Astronautica* 113 (2015), pp. 66–79. ISSN: 0094-5765. DOI: https://doi.org/10.1016/j.actaastro.2015.03.028. URL: https://www.sciencedirect.com/science/article/pii/S0094576515001289.

[4] Jeff Foust. *Starlink satellites encounter Russian ASAT debris squalls*. Ed. by Space News. https://spacenews.com/starlink-satellites-encounter-russian-asat-debris-squalls/. 2022.

[5] Eelco Doornbos. "Producing Density and Crosswind Data from Satellite Dynamics Observations". In: *Thermospheric Density and Wind Determination from Satellite Dynamics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 91–126. ISBN: 978-3-642-25129-0. DOI: 10.1007/978-3-642-25129-0_4. URL: https://doi.org/10.1007/978-3-642-25129-0_4.

[6] Bruce Bowman et al. "A New Empirical Thermospheric Density Model JB2008 Using New Solar and Geomagnetic Indices". In: *AIAA/AAS Astrodynamics Specialist Conference*. AIAA 2008-6438, 2008. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2008-6438.

[7] Raymond G. Roble. "Energetics of the Mesosphere and Thermosphere". In: *The Upper Mesosphere and Lower Thermosphere: A Review of Experiment and Theory*. American Geophysical Union (AGU), 1995, pp. 1–21. ISBN: 9781118664247. DOI: https://doi.org/

10.1029/GM087p0001. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/GM087p0001.

[8] J.T. Emmert. "Thermospheric mass density: A review". In: *Advances in Space Research* 56 (June 2015). DOI: 10.1016/j.asr.2015.05.038.

[9] Liying Qian and Stanley Solomon. "Thermospheric Density: An Overview of Temporal and Spatial Variations". In: *Space Science Reviews - SPACE SCI REV* 168 (June 2011), pp. 1–27. DOI: 10.1007/s11214-011-9810-z.

[10] Piyush M. Mehta, Richard Linares, and Eric K. Sutton. "Data-Driven Inference of Thermosphere Composition During Solar Minimum Conditions". In: *Space Weather* 17.9 (2019), pp. 1364–1379. DOI: https://doi.org/10.1029/2019SW002264. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019SW002264.

[11] L. Qian and S. Solomon. "Thermospheric Density: An Overview of Temporal and Spatial Variations". In: *Space Science Reviews* 168 (2012), pp. 147–173.

[12] Eric K. Sutton. "Interhemispheric transport of light neutral species in the thermosphere". In: *Geophysical Research Letters* 43.24 (2016), pp. 12, 325–12, 332. DOI: https://doi.org/10.1002/2016GL071679. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL071679.

[13] J. T. Houghton. "Absorption and emission by carbon-dioxide in the mesosphere". In: *Quarterly Journal of the Royal Meteorological Society* 96.410 (1970), pp. 767–770. DOI: https://doi.org/10.1002/qj.49709641025. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49709641025.

[14] G. Kockarts. "Nitric oxide cooling in the terrestrial thermosphere". In: *Geophysical Research Letters* 7.2 (1980), pp. 137–140. DOI: https://doi.org/10.1029/GL007i002p00137. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/GL007i002p00137.

[15] Marty Mlynczak et al. “The natural thermostat of nitric oxide emission at 5.3 μm in the thermosphere observed during the solar storms of April 2002”. In: *Geophysical Research Letters* 30.21 (2003). DOI: https://doi.org/10.1029/2003GL017693. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003GL017693.

[16] Sean Bruinsma et al. “Thermosphere density response to the 20–21 November 2003 solar and geomagnetic storm from CHAMP and GRACE accelerometer data”. In: *Journal of Geophysical Research: Space Physics* 111.A6 (2006). DOI: https://doi.org/10.1029/2005JA011284. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JA011284.

[17] Sean Bruinsma et al. “Thermosphere modeling capabilities assessment: geomagnetic storms”. In: *J. Space Weather Space Clim.* 11 (2021), p. 12. DOI: 10.1051/swsc/2021002. URL: https://doi.org/10.1051/swsc/2021002.

[18] J. M. T. Thompson and S. M. Tobias. “The solar dynamo”. In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 360.1801 (2002), pp. 2741–2756. DOI: 10.1098/rsta.2002.1090. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2002.1090.

[19] G. Guerrero et al. “On the Role of Tachoclines in Solar and Stellar Dynamos”. In: *The Astrophysical Journal* 819.2 (Mar. 2016), p. 104. DOI: 10.3847/0004-637x/819/2/104. URL: https://doi.org/10.3847/0004-637x/819/2/104.

[20] Gordon J. D. Petrie. “Solar Magnetism in the Polar Regions”. In: *Living Reviews in Solar Physics* 12.1 (2015). DOI: 10.1007/lrsp-2015-5. URL: https://doi.org/10.1007/lrsp-2015-5.

[21] Harold D Babcock. “The Sun’s Polar Magnetic Field”. In: *The Astrophysical Journal* 130 (1959), p. 364.

[22] E. N. Parker. “Dynamics of the Interplanetary Gas and Magnetic Fields.” In: *Astrophysical Journal* 128 (Nov. 1958), p. 664. DOI: 10.1086/146579.

[23] T. L. Garrard et al. “The Advanced Composition Explorer Mission”. In: *International Cosmic Ray Conference*. Vol. 1. International Cosmic Ray Conference. Jan. 1997, p. 105.

[24] Alexander Marshak et al. “Earth Observations from DSCOVR EPIC Instrument”. In: *Bulletin of the American Meteorological Society* 99.9 (Sept. 2018), pp. 1829–1850. DOI: 10.1175/BAMS-D-17-0223.1.

[25] Nat Gopalswamy. “Properties of Interplanetary Coronal Mass Ejections”. In: *Space Science Reviews* 124 (2006), pp. 145–168. DOI: 10.1007/s11214-006-9102-1.

[26] W. B. Manchester IV et al. “Coronal Mass Ejection Shock and Sheath Structures Relevant to Particle Acceleration”. In: *Astrophysical Journal* 622.2 (Apr. 2005), pp. 1225–1239. DOI: 10.1086/427768.

[27] Steven R. Cranmer. “Coronal Holes and the High-Speed Solar Wind”. In: *Space Science Reviews* 101.3 (2002), pp. 229–294. DOI: 10.1023/A:1020840004535.

[28] Ian G. Richardson. “Solar wind stream interaction regions throughout the heliosphere”. In: *Living Reviews in Solar Physics* 15.1 (2018). DOI: 10.1007/s41116-017-0011-z.

[29] Joseph E. Borovsky. “Solar Wind-Magnetosphere Interaction”. In: *Space Weather Fundamentals*. Ed. by George V. Khazanov. CRC Press, 2016. Chap. 4, pp. 47–73.

[30] J W Dungey. “Interplanetary Magnetic Field and the Auroral Zones”. In: *Phys. Rev. Letters* 6 (1961). DOI: 10.1103/PhysRevLett.6.47. URL: https://www.osti.gov/biblio/4087936.

[31] Homa Karimabadi et al. “Magnetic Reconnection”. In: *Space Weather Fundamentals*. Ed. by George V. Khazanov. CRC Press, 2016. Chap. 6, pp. 95–113.

[32] D. H. Fairfield and L. J. Cahill Jr. “Transition region magnetic field and polar magnetic disturbances”. In: *Journal of Geophysical Research (1896-1977)* 71.1 (1966), pp. 155–169. DOI: https://doi.org/10.1029/JZ071i001p00155. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ071i001p00155.

[33] Jeffrey M. Forbes. “Dynamics of the Thermosphere”. In: *Journal of the Meteorological Society of Japan. Ser. II* 85B (2007), pp. 193–213. DOI: 10.2151/jmsj.85B.193.

[34] G. R. Wilson et al. “Response of the thermosphere to Joule heating and particle precipitation”. In: *Journal of Geophysical Research: Space Physics* 111.A10 (2006). DOI: https://doi.org/10.1029/2005JA011274. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JA011274.

[35] G. W. Prölss et al. “Ionospheric storm effects at subauroral latitudes: A case study”. In: *Journal of Geophysical Research* 96.A2 (Feb. 1991), pp. 1275–1288. DOI: 10.1029/90JA02326.

[36] Sean L. Bruinsma and Jeffrey M. Forbes. “Global observation of traveling atmospheric disturbances (TADs) in the thermosphere”. In: *Geophysical Research Letters* 34.14 (2007). DOI: https://doi.org/10.1029/2007GL030243. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007GL030243.

[37] Martin Asplund et al. “The Chemical Composition of the Sun”. In: *Annual Review of Astronomy and Astrophysics* 47.1 (2009), pp. 481–522. DOI: 10.1146/annurev.astro.46.060407.145222. URL: https://doi.org/10.1146/annurev.astro.46.060407.145222.

[38] Gábor Tóth et al. “Space Weather Modeling Framework: A new tool for the space science community”. In: *Journal of Geophysical Research: Space Physics* 110.A12 (2005). DOI: https://doi.org/10.1029/2005JA011126. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JA011126.

[39] J. M. Picone et al. “NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues”. In: *Journal of Geophysical Research: Space Physics* 107.A12 (2002), SIA 15-1-SIA 15–16. DOI: 10.1029/2002JA009430. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JA009430.

[40] Bruinsma, Sean. “The DTM-2013 thermosphere model”. In: *J. Space Weather Space Clim.* 5 (2015), A1. DOI: 10.1051/swsc/2015001. URL: https://doi.org/10.1051/swsc/2015001.

[41] Mark Storz, Bruce Bowman, and James Branson. "High Accuracy Satellite Drag Model (HASDM)". In: *AIAA/AAS Astrodynamics Specialist Conference and Exhibit*. 2005. DOI: 10.2514/6.2002-4886. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2002-4886.

[42] H. Luhr, L. Grunwaldt, and C. Forste. *CHAMP Reference Systems, Transformations and Standards*. Tech. rep. GFZ-Potsdam, Postdam, Germany. CH-GFZ-RS-002, 2002.

[43] S. Bettadpur. "Gravity Recovery and Climate Experiment: Product Specification Document". In: *GRACE 327-720, CSR-GR-03-02* (2012). Cent. for Space Res., The Univ. of Texas, Austin, TX.

[44] Sean Bruinsma and Richard Biancale. "Total Densities Derived from Accelerometer Data". In: *Journal of Spacecraft and Rockets* 40.2 (2003), pp. 230–236. DOI: 10.2514/2.3937. eprint: https://doi.org/10.2514/2.3937. URL: https://doi.org/10.2514/2.3937.

[45] H. Liu et al. "Global distribution of the thermospheric total mass density derived from CHAMP". In: *Journal of Geophysical Research: Space Physics* 110.A4 (2005). DOI: https://doi.org/10.1029/2004JA010741. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010741.

[46] Eric K. Sutton. "Effects of solar disturbances on the thermosphere densities and winds from CHAMP and GRACE satellite accelerometer data". PhD thesis. University of Colorado at Boulder, Oct. 2008. URL: https://ui.adsabs.harvard.edu/abs/2008PhDT.....87S.

[47] Andres Calabia and Shuanggen Jin. "New modes and mechanisms of thermospheric mass density variations from GRACE accelerometers". In: *Journal of Geophysical Research: Space Physics* 121.11 (2016), pp. 11, 191–11, 212. DOI: https://doi.org/10.1002/2016JA022594.

[48] Piyush M. Mehta et al. "New density estimates derived using accelerometers on board the CHAMP and GRACE satellites". In: *Space Weather* 15.4 (2017), pp. 558–576. DOI: https://doi.org/10.1002/2016SW001562. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016SW001562.

[49] Liying Qian et al. “The NCAR TIE-GCM: A community model of the coupled thermosphere/ionosphere system”. In: *Geophysical Monograph Series* 201 (Jan. 2013), pp. 73–83. DOI: 10.1029/2012GM001297.

[50] T. J. Fuller-Rowell et al. “Storm-time changes in the upper atmosphere at low latitudes”. In: *Journal of Atmospheric and Solar-Terrestrial Physics* 64.12-14 (Aug. 2002), pp. 1383–1391. DOI: 10.1016/S1364-6826(02)00101-3.

[51] A.J. Ridley, Y. Deng, and G. Tóth. “The global ionosphere–thermosphere model”. In: *Journal of Atmospheric and Solar-Terrestrial Physics* 68.8 (2006), pp. 839–864. ISSN: 1364-6826. DOI: https://doi.org/10.1016/j.jastp.2006.01.008. URL: http://www.sciencedirect.com/science/article/pii/S1364682606000071.

[52] J. T. Emmert et al. “NRLMSIS 2.0: A Whole-Atmosphere Empirical Model of Temperature and Neutral Species Densities”. In: *Earth and Space Science* 8.3 (2021), e2020EA001321. DOI: https://doi.org/10.1029/2020EA001321. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020EA001321.

[53] Alan E. Hedin. “MSIS-86 Thermospheric Model”. In: *Journal of Geophysical Research: Space Physics* 92.A5 (1987), pp. 4649–4662. DOI: https://doi.org/10.1029/JA092iA05p04649. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA092iA05p04649.

[54] F. Marcos et al. “Precision Low Earth Orbit Determination Using Atmospheric Density Calibration”. In: *Journal of The Astronautical Sciences* 46 (1998), pp. 395–409.

[55] A.I. Nazarenko, P.J. Cefola, and V. Yurasov. “Estimating atmospheric density variations to improve LEO orbit prediction accuracy”. In: *AIAA/AAS Space Flight Mechanics Meeting*. AAS 98-190, 1998.

[56] B. Bowman and M. Storz. “High Accuracy Satellite Drag Model (HASDM) Review”. In: *AIAA/AAS Astrodynamics Specialist Conference*. AAS 03-625, 2003.

[57] W. Kent Tobiska et al. “The SET HASDM density database”. In: *Space Weather* (2021). DOI: https://doi.org/10.1029/2020SW002682.

[58] Robert E. Dickinson, E. C. Ridley, and R. G. Roble. "A three-dimensional general circulation model of the thermosphere". In: *Journal of Geophysical Research: Space Physics* 86.A3 (1981), pp. 1499–1512. DOI: https://doi.org/10.1029/JA086iA03p01499. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA086iA03p01499.

[59] A. D. Richmond, E. C. Ridley, and R. G. Roble. "A thermosphere/ionosphere general circulation model with coupled electrodynamics". In: *Geophysical Research Letters* 19.6 (1992), pp. 601–604. DOI: https://doi.org/10.1029/92GL00401. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/92GL00401.

[60] Piyush M. Mehta, Craig A. McLaughlin, and Eric K. Sutton. "Drag coefficient modeling for grace using Direct Simulation Monte Carlo". In: *Advances in Space Research* 52.12 (2013), pp. 2035–2051. ISSN: 0273-1177. DOI: https://doi.org/10.1016/j.asr.2013.08.033. URL: https://www.sciencedirect.com/science/article/pii/S0273117713005486.

[61] Piyush M. Mehta et al. "Comparing Physical Drag Coefficients Computed Using Different Gas–Surface Interaction Models". In: *Journal of Spacecraft and Rockets* 51.3 (2014), pp. 873–883. DOI: 10.2514/1.A32566. URL: https://doi.org/10.2514/1.A32566.

[62] Andrew Walker, Piyush Mehta, and Josef Koller. "Drag Coefficient Model Using the Cercignani–Lampis–Lord Gas–Surface Interaction Model". In: *Journal of Spacecraft and Rockets* 51.5 (2014), pp. 1544–1563. DOI: 10.2514/1.A32677. URL: https://doi.org/10.2514/1.A32677.

[63] D. R. Bates. "Some Problems concerning the Terrestrial Atmosphere above about the 100 km Level". In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 253.1275 (1959), pp. 451–462. ISSN: 00804630. URL: http://www.jstor.org/stable/100692.

[64] Luigi G. Jacchia. "Static Diffusion Models of the Upper Atmosphere with Empirical Temperature Profiles". In: *Smithsonian Contributions to Astrophysics* 8 (Jan. 1965), p. 215. URL: https://ui.adsabs.harvard.edu/abs/1965SCoA....8..215J.

[65] James C. G. Walker. “Analytic Representation of Upper Atmosphere Densities Based on Jacchia’s Static Diffusion Models”. In: *Journal of Atmospheric Sciences* 22.4 (1965), pp. 462–463. DOI: 10.1175/1520-0469(1965)022<0462:AROUAD>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/atsc/22/4/1520-0469_1965_022_0462_arouad_2_0_co_2.xml.

[66] Jeffrey M. Forbes et al. “Surface-exosphere coupling due to thermal tides”. In: *Geophysical Research Letters* 36.15 (2009). DOI: https://doi.org/10.1029/2009GL038748. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009GL038748.

[67] Jeffrey M. Forbes et al. “Sun-synchronous thermal tides in exosphere temperature from CHAMP and GRACE accelerometer measurements”. In: *Journal of Geophysical Research: Space Physics* 116.A11 (2011). DOI: https://doi.org/10.1029/2011JA016855. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JA016855.

[68] D. R. Weimer et al. “Intercalibration of neutral density measurements for mapping the thermosphere”. In: *Journal of Geophysical Research: Space Physics* 121.6 (2016), pp. 5975–5990. DOI: https://doi.org/10.1002/2016JA022691. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JA022691.

[69] D. R. Weimer et al. “Improving Neutral Density Predictions Using Exospheric Temperatures Calculated on a Geodesic, Polyhedral Grid”. In: *Space Weather* 18.1 (2020), e2019SW002355. DOI: https://doi.org/10.1029/2019SW002355. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019SW002355.

[70] Daniel R. Weimer et al. “Comparison of a Neutral Density Model With the SET HASDM Density Database”. In: *Space Weather* 19.12 (2021), e2021SW002888. DOI: https://doi.org/10.1029/2021SW002888. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002888.

[71] E. Friis-Christensen, H. Lühr, and G. Hulot. “Swarm: A constellation to study the Earth’s magnetic field”. In: *Earth, Planets and Space* 58 (2006), pp. 351–358. DOI: 10.1186/BF03351933. URL: https://doi.org/10.1186/BF03351933.

[72] E. Astafyeva et al. “Global Ionospheric and Thermospheric Effects of the June 2015 Geomagnetic Disturbances: Multi-Instrumental Observations and Modeling”. In: *Journal of Geophysical Research: Space Physics* 122.11 (2017), pp. 11, 716–11, 742. DOI: https://doi.org/10.1002/2017JA024174. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JA024174.

[73] W. Kent Tobiska, Bruce Bowman, and S. Dave Bouwer. “Solar and Geomagnetic Indices for the JB2008 Thermosphere Density Model”. In: COSPAR CIRA Draft, 2008. Chap. 4.

[74] ISO 21348. *Space environment (natural and artificial) -Process for determining solar irradiances*. Standard. Geneva, CH: International Organization for Standardization, May 2007.

[75] D.A. Vallado and W.D. McClain. *Fundamentals of Astrodynamics and Applications*. Space Technology Library. Springer Netherlands, 2001, pp. 556–557. ISBN: 9780792369035. URL: https://books.google.com/books?id=PJLlWzMBKjkC.

[76] ISO 14222. *Space environment (natural and artificial) -Earth upper atmosphere*. Standard. Geneva, CH: International Organization for Standardization, Mar. 2013.

[77] D. R. Weimer. “Improved ionospheric electrodynamic models and application to calculating Joule heating rates”. In: *Journal of Geophysical Research: Space Physics* 110.A5 (2005). DOI: https://doi.org/10.1029/2004JA010884. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010884.

[78] D. R. Weimer. “Predicting surface geomagnetic variations using ionospheric electrodynamic models”. In: *Journal of Geophysical Research: Space Physics* 110.A12 (2005). DOI: https://doi.org/10.1029/2005JA011270. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JA011270.

[79] Toshihiko Iyemori. “Storm-Time Magnetospheric Currents Inferred from Mid-Latitude Geomagnetic Field Variations”. In: *Journal of geomagnetism and geoelectricity* 42.11 (1990), pp. 1249–1265. DOI: 10.5636/jgg.42.1249.

[80] Warren S. McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The Bulletin of Mathematical Biophysics* 5 (1943), pp. 115–133. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259.

[81] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), pp. 386–408.

[82] Seppo Linnainmaa. “The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors”. MA thesis. University of Helsinki, 1970.

[83] S. Dreyfus. “The computational solution of optimal control problems with time lag”. In: *IEEE Transactions on Automatic Control* 18.4 (1973), pp. 383–385. DOI: 10.1109/TAC.1973.1100330.

[84] Kumar Chellapilla, Sidd Puri, and Patrice Simard. “High Performance Convolutional Neural Networks for Document Processing”. In: *Tenth International Workshop on Frontiers in Handwriting Recognition*. Ed. by Guy Lorette. Université de Rennes 1. La Baule (France): Suvisoft, 2006. URL: https://hal.inria.fr/inria-00112631.

[85] Kaiming He et al. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015. DOI: 10.48550/ARXIV.1502.01852. URL: https://arxiv.org/abs/1502.01852.

[86] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. DOI: 10.48550/ARXIV.1412.6980. URL: https://arxiv.org/abs/1412.6980.

[87] Timothy Dozat. “Incorporating Nesterov Momentum into Adam”. In: *ICLR 2016 Workshop*. 2016.

[88] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Journal of Machine Learning Research* 12.61 (2011), pp. 2121–2159. URL: http://jmlr.org/papers/v12/duchi11a.html.

[89] Matthew D. Zeiler. *ADADELTA: An Adaptive Learning Rate Method*. 2012. DOI: 10.48550/ARXIV.1212.5701. URL: https://arxiv.org/abs/1212.5701.

[90] Alex Graves. *Generating Sequences With Recurrent Neural Networks*. 2013. DOI: 10.48550/ARXIV.1308.0850. URL: https://arxiv.org/abs/1308.0850.

[91] J. Sola and J. Sevilla. "Importance of input data normalization for the application of neural networks to complex industrial problems". In: *IEEE Transactions on Nuclear Science* 44.3 (1997), pp. 1464–1468. DOI: 10.1109/23.589532.

[92] Tom O'Malley et al. *Keras Tuner*. https://github.com/keras-team/keras-tuner. 2019.

[93] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[94] Karl Pearson F.R.S. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720. eprint: https://doi.org/10.1080/14786440109462720. URL: https://doi.org/10.1080/14786440109462720.

[95] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441. DOI: 10.1037/h0071325.

[96] H Bjornsson and S. A. Venegas. *A manual for EOF and SVD analyses of climatic data*. Tech. rep. Report No. 97-1, Montreal, Quebec. MCGill Univ., 1997.

[97] Carl Edward Rasmussen. "Gaussian Processes in Machine Learning". In: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Ed. by Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch. Berlin, Heidelberg: Springer Berlin Heidel-

berg, 2004, pp. 63–71. ISBN: 978-3-540-28650-9. DOI: 10.1007/978-3-540-28650-9_4. URL: https://doi.org/10.1007/978-3-540-28650-9_4.

[98] Jie Wang. *An Intuitive Tutorial to Gaussian Processes Regression*. 2021. arXiv: 2009.10862 [stat.ML].

[99] M. Chandorkar, E. Camporeale, and S. Wing. "Probabilistic forecasting of the disturbance storm time index: An autoregressive Gaussian process approach". In: *Space Weather* 15.8 (2017), pp. 1004–1019. DOI: https://doi.org/10.1002/2017SW001627. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017SW001627.

[100] Tianyu Gao, Hao Peng, and Xiaoli Bai. "Calibration of atmospheric density model based on Gaussian Processes". In: *Acta Astronautica* 168 (2020), pp. 273–281. ISSN: 0094-5765. DOI: https://doi.org/10.1016/j.actaastro.2019.12.014. URL: https://www.sciencedirect.com/science/article/pii/S009457651931447X.

[101] Haitao Liu, Jianfei Cai, and Yew-Soon Ong. "Remarks on multi-output Gaussian process regression". In: *Knowledge-Based Systems* 144 (2018), pp. 102–121. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2017.12.034. URL: https://www.sciencedirect.com/science/article/pii/S0950705117306123.

[102] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (June 2014), pp. 1929–1958.

[103] Richard Alake. *Understanding and Implementing Dropout in TensorFlow and Keras*. Ed. by towards data science. https://towardsdatascience.com/understanding-and-implementing-dropout-in-tensorflow-and-keras-a8a3a02c1bfa. May 2020.

[104] Yarin Gal and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2016. arXiv: 1506.02142 [stat.ML].

[105] Piyush M. Mehta. "Thermospheric density and satellite drag modeling". PhD thesis. University of Kansas, Jan. 2013. URL: https://ui.adsabs.harvard.edu/abs/2013PhDT.....90M.

[106] B. Efron. “Bootstrap Methods: Another Look at the Jackknife”. In: *The Annals of Statistics* 7.1 (1979), pp. 1–26. ISSN: 00905364. URL: http://www.jstor.org/stable/2958830.

[107] James E. Matheson and Robert L. Winkler. “Scoring Rules for Continuous Probability Distributions”. In: *Management Science* 22.10 (1976), pp. 1087–1096. ISSN: 00251909, 15265501. URL: http://www.jstor.org/stable/2629907.

[108] Enrico Camporeale and Algo Care. “ACCRUE: Accurate and Reliable Uncertainty Estimate in Deterministic Models”. In: *International Journal for Uncertainty Quantification* 11.4 (2021), pp. 81–94. ISSN: 2152-5080.

[109] Tilmann Gneiting et al. “Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation”. In: *Monthly Weather Review* 133.5 (2005), pp. 1098–1118. DOI: 10.1175/MWR2904.1. URL: https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2904.1.xml.

[110] Gemma J. Anderson et al. *Meaningful uncertainties from deep neural network surrogates of large-scale numerical simulations*. 2020. arXiv: 2010.13749 [stat.ML].

[111] A.C. Davison et al. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997, pp. 243–244. ISBN: 9780521574716.

[112] Darko Pevec and Igor Kononenko. “Prediction intervals in supervised learning for model evaluation and discrimination”. In: *Applied Intelligence* 42 (2014), pp. 790–804.

[113] Richard J. Licata et al. “Qualitative and Quantitative Assessment of the SET HASDM Database”. In: *Space Weather* 19.8 (2021). DOI: https://doi.org/10.1029/2021SW002798. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002798.

[114] D.A. Nix and A.S. Weigend. “Estimating the mean and variance of the target probability distribution”. In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94)*. Vol. 1. 1994, pp. 55–60. DOI: 10.1109/ICNN.1994.374138.

[115] Richard J. Licata et al. “Improved Neutral Density Predictions Through Machine Learning Enabled Exospheric Temperature Model”. In: *Space Weather* 19.12 (2021), e2021SW002918. DOI: https://doi.org/10.1029/2021SW002918. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021SW002918. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002918.

[116] Piyush M. Mehta, Richard Linares, and Eric K. Sutton. “A Quasi-Physical Dynamic Reduced Order Model for Thermospheric Mass Density via Hermitian Space-Dynamic Mode Decomposition”. In: *Space Weather* 16.5 (2018), pp. 569–588. DOI: 10.1029/2018SW001840. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW001840.

[117] David J. Gondelach and Richard Linares. “Real-Time Thermospheric Density Estimation via Two-Line Element Data Assimilation”. In: *Space Weather* 18.2 (2020), e2019SW002356. DOI: https://doi.org/10.1029/2019SW002356. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019SW002356.

[118] David J. Gondelach and Richard Linares. “Real-Time Thermospheric Density Estimation via Radar and GPS Tracking Data Assimilation”. In: *Space Weather* 19.4 (2021), e2020SW002620. DOI: https://doi.org/10.1029/2020SW002620. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020SW002620. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020SW002620.

[119] Joshua L. Proctor, Steven L. Brunton, and J. Nathan Kutz. “Dynamic Mode Decomposition with Control”. In: *SIAM Journal on Applied Dynamical Systems* 15.1 (2016), pp. 142–161. DOI: 10.1137/15M1013857. eprint: https://doi.org/10.1137/15M1013857. URL: https://doi.org/10.1137/15M1013857.

[120] S. B. Xu et al. “Prediction of the Dst Index with Bagging Ensemble-learning Algorithm”. In: *The Astrophysical Journal Supplement Series* 248.1 (May 2020), p. 14. DOI: 10.3847/1538-4365/ab880e. URL: https://doi.org/10.3847/1538-4365/ab880e.

[121] Peian Wang et al. “The Prediction of Storm-Time Thermospheric Mass Density by LSTM-Based Ensemble Learning”. In: *Space Weather* 20.3 (2022). e2021SW002950 2021SW002950, e2021SW002950. DOI: https://doi.org/10.1029/2021SW002950. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021SW002950. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002950.

[122] Max-Heinrich Laves et al. “Recalibration of Aleatoric and EpistemicRegression Uncertainty in Medical Imaging”. In: *Machine Learning for Biomedical Imaging* 1 (MIDL 2020 special issue 2021). ISSN: 2766-905X. URL: https://melba-journal.org/papers/2021:008.html.

[123] Kostas Florios et al. “Forecasting solar flares using magnetogram-based predictors and machine learning”. In: *Solar Physics* 293.2 (2018), pp. 1–42.

[124] Yu Jiao, John J. Hall, and Yu T. Morton. “Automatic Equatorial GPS Amplitude Scintillation Detection Using a Machine Learning Algorithm”. In: *IEEE Transactions on Aerospace and Electronic Systems* 53.1 (2017), pp. 405–418. DOI: 10.1109/TAES.2017.2650758.

[125] S. Wing et al. “Kp forecast models”. In: *Journal of Geophysical Research: Space Physics* 110.A4 (2005). DOI: https://doi.org/10.1029/2004JA010500. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010500.

[126] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[127] Eftyhia Zesta and Denny M. Oliveira. “Thermospheric Heating and Cooling Times During Geomagnetic Storms, Including Extreme Events”. In: *Geophysical Research Letters* 46.22 (2019), pp. 12739–12746. DOI: https://doi.org/10.1029/2019GL085120. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085120.

[128] J. M. III Russell et al. “An Overview of the SABER Experiment and Preliminary Calibration Results”. In: *Space Dynamics Lab Publications* 114 (1999). DOI: 10.1117/12.366382. URL: https://digitalcommons.usu.edu/sdl_pubs/114.

[129] Jiuhou Lei et al. “Overcooling in the upper thermosphere during the recovery phase of the 2003 October storms”. In: *Journal of Geophysical Research: Space Physics* 117.A3 (2012). DOI: https://doi.org/10.1029/2011JA016994. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JA016994.

[130] D. J. Knipp et al. “Thermospheric nitric oxide response to shock-led storms”. In: *Space Weather* 15.2 (2017), pp. 325–342. DOI: https://doi.org/10.1002/2016SW001567. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016SW001567.

[131] Andrey V. Mikhailov and Loredana Perrone. “Poststorm Thermospheric NO Overcooling?” In: *Journal of Geophysical Research: Space Physics* 125.1 (2020), e2019JA027122. DOI: https://doi.org/10.1029/2019JA027122. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JA027122.

[132] Jiuhou Lei et al. “Comments on “Poststorm Thermospheric NO Overcooling?” by Mikhailov and Perrone (2020)”. In: *Journal of Geophysical Research: Space Physics* 126.4 (2021), e2020JA027992. DOI: https://doi.org/10.1029/2020JA027992. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JA027992.

[133] D. M. Oliveira and E. Zesta. “Satellite Orbital Drag During Magnetic Storms”. In: *Space Weather* 17.11 (2019), pp. 1510–1533. DOI: https://doi.org/10.1029/2019SW002287. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019SW002287.

[134] Denny M. Oliveira et al. “The Current State and Future Directions of Modeling Thermosphere Density Enhancements During Extreme Magnetic Storms”. In: *Frontiers in Astronomy and Space Sciences* 8 (2021). ISSN: 2296-987X. DOI: 10.3389/fspas.2021.764144.

[135] Charles D. Bussy-Virat, Aaron J. Ridley, and Joel W. Getchius. “Effects of Uncertainties in the Atmospheric Density on the Probability of Collision Between Space Objects”. In: *Space Weather* 16.5 (2018), pp. 519–537. DOI: 10.1029/2017SW001705.

[136] W.Kent Tobiska et al. “The SOLAR2000 empirical solar irradiance model and forecast tool”. In: *Journal of Atmospheric and Solar-Terrestrial Physics* 62.14 (2000), pp. 1233–1250. ISSN: 1364-6826. DOI: https://doi.org/10.1016/S1364-6826(00)00070-5.

[137] W. Kent Tobiska, S. Dave Bouwer, and Bruce R. Bowman. “The development of new solar indices for use in thermospheric density modeling”. In: *Journal of Atmospheric and Solar-Terrestrial Physics* 70.5 (2008), pp. 803–819. ISSN: 1364-6826. DOI: https://doi.org/10.1016/j.jastp.2007.11.001.

[138] Richard J. Licata, Piyush M. Mehta, and Christina Kay. “Data-Driven Framework for Space Weather Modeling with Uncertainty Treatment towards Space Situational Awareness and Space Traffic Management”. In: *Astrodynamics Specialist Conference* (Portland, ME). AAS 19-603, Aug. 2019.

[139] S.N. Paul et al. “SATELLITE DRAG COEFFICIENT MODELING AND ORBIT UNCERTAINTY QUANTIFICATION USING STOCHASTIC MACHINE LEARNING TECHNIQUES”. In: *2021 AAS/AIAA Astrodynamics Specialist Conference* (2021). URL: https://par.nsf.gov/biblio/10315463.

[140] Craig A. McLaughlin et al. “Estimating Density Using Precision Satellite Orbits from Multiple Satellites”. In: *The Journal of the Astronautical Sciences* 59 (2012), pp. 84–100. DOI: 10.1007/s40295-013-0007-4.

[141] Bob Schutz, Byron Tapley, and George Born. “Statistical Orbit Determination”. In: 1st ed. Burlington, MA: Elsevier Academic Press, 2004. Chap. 6.

[142] Carmen Pardini and Luciano Anselmo. “Revisiting the collision risk with cataloged objects for the Iridium and COSMO-SkyMed satellite constellations”. In: *Acta Astronautica* 134

(2017), pp. 23–32. ISSN: 0094-5765. DOI: https://doi.org/10.1016/j.actaastro.2017.01.046. URL: https://www.sciencedirect.com/science/article/pii/S0094576516312607.

[143] Aaron C. Boley and Michael Byers. “Satellite mega-constellations create risks in Low Earth Orbit, the atmosphere and on Earth”. In: *Scientific Reports* 11 (2021). ISSN: 2045-2322. DOI: 10.1038/s41598-021-89909-7.