SCIENCE THROUGH MACHINE LEARNING: A CASE STUDY FOR POSTSTORM THERMOSPHERIC COOLING

Richard J. Licata
Dept. of Mechanical and Aerospace Engineering
West Virginia University
Morgantown, WV 26505
rjlicata@mix.wvu.edu

Piyush M. Mehta
Dept. of Mechanical and Aerospace Engineering
West Virginia University
Morgantown, WV

Daniel R. Weimer
Center for Space Science and Engineering Research
Virginia Tech
Blacksburg, VA

Douglas P. Doug
Space Science Division
U.S. Naval Research Laboratory
Washington, DC

W. Kent Tobiska
Space Environments Technologies
Pacific Palisades, CA

Jean Yoshii
Space Environments Technologies
Pacific Palisades, CA

February 17, 2022

ABSTRACT

Machine learning (ML) is often viewed as a black-box regression technique that is typically unable to provide considerable scientific value or insight. ML models are valuable universal function approximators and – if used correctly – can provide scientific information related to the ground-truth dataset used for fitting. A benefit to ML over parametric models is that there are no basis function limiting what phenomena it can represent. In this work, we develop ML models on three datasets: the Space Environment Technologies (SET) High Accuracy Satellite Drag Model (HASDM) density database, a spatiotemporally matched dataset of outputs from the Jacchia-Bowman 2008 Empirical Thermospheric Density Model (JB2008), and an accelerometer-derived density dataset from CHAllenging Minisatellite Payload (CHAMP). These ML models are compared to the Naval Research Laboratory Mass Spectrometer and Incoherent Scatter radar (NRLMSIS 2.0) model to study the presence of post-storm cooling in the middle-thermosphere. We find that both NRLMSIS 2.0 and JB2008-ML do not account for post-storm cooling and subsequently perform poorly in periods following strong geomagnetic storms (e.g. the 2003 Halloween storms). Conversely, HASDM-ML and CHAMP-ML do show evidence of post-storm cooling indicating that this phenomenon is present in the original datasets.

1 Introduction

Machine learning (ML) has been becoming increasingly prevalent across many scientific domains over the past several decades. This has been aided largely by the improvements in computer processing power and the implementation of Graphical Processing Units (GPUs) in model training [1]. As it pertains to the space weather community, ML has been used to develop models for problems such as solar flare prediction [2], ionospheric scintillation detection [3], and geomagnetic index forecasts [4]. However, its use is often limited to problem solving, not for investigative purposes. Furthermore, there are rarely any studies on what the model has learned outside of determining its performance metrics. Convolutional models – specifically related to image processing – are inherently easier to understand, as the weights or

filters can be displayed as images [5]. This is not a luxury associated with ML regression models which motivates this work as it pertains to space weather and the thermosphere.

The thermosphere is the neutral region of the upper atmosphere consisting of many atomic and molecular species. Their relative abundance and contribution to the total mass density can vary as a function of altitude, solar activity, and space weather conditions [6]. During abrupt space weather events like geomagnetic storms, the local distributions of the constituents can vary considerably causing significant changes in the total mass density [7]. Zesta and Oliveira (2019) found that when storms become stronger, the thermosphere both heats and cools at a faster rate [8]. Significant research has been done into a potential cause of the cooling effects, overproduction of nitric oxide (NO) and its infrared emissions.

Kockarts (1980) 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 [9]. 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 [10] and high fidelity density estimates from satellite such as CHAllenging Minisatellite Payload (CHAMP) [11] and Gravity Recovery and Climate Experiment (GRACE) [12].

Mlynczak et al. (2003) used SABER data during the storm periods of April 2002 and found that NO emissions were notably enhanced during this period [13]. Lei et al (2012) 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 [14]. Knipp et al. (2017) 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 [15]. As this is an active research area, the driving force behind the cooling effect is still disputed [16, 17] and 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 model development process and how we use them to examine this phenomena. 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.

2 Data, Models, and Methods

2.1 Data and Models

As a benchmark, we use the Naval Research Laboratory Mass Spectrometer and Incoherent Scatter radar (NRLMSIS 2.0) empirical thermosphere model [18]. This is the most recent version of MSIS models that date back to the original MSIS-86 model [19]. NRLMSIS 2.0 uses the *ap* index to account for geomagnetic activity. The *ap* index is indicative of global geomagnetic activity and has a three hour cadence. 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 manuscript.

We also develop machine-learned models (to be described in Section 2.2) based on three datasets. The first is outputs of the Jacchia-Bowman 2008 Empirical Thermospheric Density Model (JB2008) from the start of 2000 to end the of 2019 [20]. We evaluated JB2008 every three hours and at a fixed grid of 12,312 locations. The longitude, latitude, and altitude resolutions are 15°, 10°, and 25 km, respectively with the altitude ranging from 175 – 825 km. JB2008 is driven by four solar indices/proxies. As with NRLMSIS 2.0, it uses $F_{10.7}$ as a driver for solar activity. $F_{10.7}$ is a valuable solar proxy that represents the 10.7 cm solar radio emission [21]. JB2008 also uses $S_{10.7}$, $M_{10.7}$, and $Y_{10.7}$, which are explained by Tobiska et al. (2008) and Bowman et al. (2008) [22, 20]. We do not go into their details in this work as it is not pertinent to this analysis. For geomagnetic activity, JB2008 uses both *ap* and *Dst*. *Dst* is an index for the strength of the ring current which serves as a useful proxy for geomagnetic activity.

The second dataset upon which we develop a ML model is the Space Environment Technologies (SET) High Accuracy Satellite Drag Model (HASDM) density database [23]. This was the first major release of outputs from the U.S. Air Force's HASDM model. This HASDM dataset is spatiotemporally matched to the JB2008 outputs used here (same time period, time resolution, and locations). HASDM is often considered the state-of-the-art for thermospheric density

modeling as it assimilates observed drag data from calibration satellites to make corrections to its background empirical model, JB2008 [24].

The final dataset considered is the Mehta et al. (2017) accelerometer-derived density estimates from CHAMP [25]. CHAMP was in orbit from 2000 – 2010 with a high inclination and altitude range of 300 – 460 km. Unlike JB2008 and HASDM, this is an in-situ dataset with a much higher cadence – 10 seconds. There are numerous other datasets that have been developed using CHAMP accelerometer data [26, 27, 28, 29, 30], but we proceed with the described dataset due to its use in previous work [31].

2.2 Machine Learning

As the goal is to study the effects of geomagnetic time histories, we do not need to develop a surrogate model for NRLMSIS 2.0 (see Section 2.1). We do, however, proceed with model development for the JB2008, HASDM, and CHAMP datasets. The process for JB2008 and HASDM is identical as they have the same time and space resolution. In addition, HASDM is rooted in JB2008, so we use the same inputs for both datasets. To keep model size reasonable, we leverage principal component analysis (PCA) to reduce the spatial dimensionality from 12,312 to 10 PCA coefficients. This has been demonstrated on similar datasets in the past [32, 33, 34]. For information on the data preparation and PCA, the reader is referred to Licata et al. (2021) [31]. The major difference for CHAMP is that it is an in-situ dataset, so location is now an input as opposed to being embedded in the model output.

We first prepare the data for ML, defining the input and output data structures. JB2008 and HASDM have 26 inputs, defined in Table 1. Eight of the inputs are the four solar drivers described in Section 2.1 along with their 81-day centered averages – marked with an "81c" subscript. These inputs are also shared with the CHAMP model. For geomagnetic activity, both JB2008 and HASDM use time histories for *ap* and *Dst*. The *ap* time series is the one described for NRLMSIS 2.0, and the *Dst* time series was determined in previous work [34]. Since CHAMP has a much higher time resolution (1 minute), we forgo the use of these geomagnetic indices and instead use *SYM-H*, which represents the longitudinally symmetric geomagnetic field disturbances [35, 36]. The time history was chosen to be similar to those used by the other models. We also use the Poynting flux totals in the northern and southern hemispheres (S_N and S_S) generated by the W05 electrodynamics model [37, 38]. The time and location inputs are described in Equations 1 and 2. The general form of $(2\pi x/y)$ allows for the linearly increasing inputs to be continuous about their boundaries. t_1 and t_2 represent annual variations, using the day of year (doy). t_3 and t_4 represent diurnal variations, using universal time (UT). *LST* in Equation 2 is the local solar time input.

$$t_1 = sin\left(\frac{2\pi doy}{365.25}\right), \quad t_2 = cos\left(\frac{2\pi doy}{365.25}\right), \quad t_3 = sin\left(\frac{2\pi UT}{24}\right), \quad t_4 = cos\left(\frac{2\pi UT}{24}\right). \tag{1}$$

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

Table 1: List of inputs for the ML models. Note: *LAT* and *ALT* are the latitude and altitude at epoch, respectively.

JB2008 / HASDM		
Solar	**Geomagnetic**	**Temporal**
F_{10}, S_{10}, M_{10}, Y_{10}, F_{81c}, S_{81c}, M_{81c}, Y_{81c}	a_{pA}, a_p, a_{p3}, a_{p6}, a_{p9}, a_{p12-33}, a_{p36-57}, 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
CHAMP		
Solar	**Geomagnetic**	**Spatial/Temporal**
F_{10}, S_{10}, M_{10}, Y_{10}, F_{81c}, S_{81c}, M_{81c}, Y_{81c}	*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*$_{33\text{-}57}$, S_N, S_S	LST_1, LST_2, LAT, ALT, t_1, t_2, t_3, t_4

The outputs for JB2008 and HASDM are their respective 10 PCA coefficients, while the CHAMP outputs are the local density estimates. With the inputs and outputs set up in an ML format, we can determine an architecture for

each dataset using Keras Tuner [39]. The tuner is provided a hyperparameter space (e.g. choices for the number of layers, neurons, activation functions, and optimizers) and trains three models with random weight initialization for 100 architectures or trials. The first 25 architectures are randomly selected from the search space. After the random search, the tuner chooses all subsequent architectures using a Bayesian optimization scheme attempting to minimize the loss on the validation set. The ten best models are evaluated on the training and validation sets to determine the final model considering the lowest errors. The model development for CHAMP, HASDM, and subsequently JB2008 is outlined by Licata and Mehta (2022) with the only difference being the *SYM-H* time series inputs for CHAMP [40]. We provide additional details on model development in Appendices A and B.

2.3 Storm Example

To motivate the work, we evaluate NRLMSIS 2.0 and the ML models from the JB2008, HASDM, and CHAMP datasets during the 2003 Halloween storms. The latter models will be referred to as JB2008-ML, HASDM-ML, and CHAMP-ML, respectively. The four models are provided the true drivers for the six day period from October 28 – November 3, 2003 and are compared to the Mehta et al. (2017) CHAMP estimates. For NRLMSIS 2.0 and CHAMP-ML, the predictions are 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, we make prediction at the 3-hour intervals used by the original models. Those 3D density grids are then interpolated in space and time in log-scale to the locations of CHAMP. The final step is to take a 92.5 minute running average of the orbit densities to obtain orbit-average densities. This allows us to visualize the general density along the orbit during this period.

2.4 Time Lag Study

As discussed in Section 1, cooling mechanisms often cause post-storm densities to be anomalously low. For this storm in particular, Lei et al. (2012) 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, we vary the time histories for *ap* or *SYM-H* independently within the models at four locations listed in Table 2. Table 2 also contains the geomagnetic indices held constant in each model while either *ap* or *SYM-H* are being changed. All cases are at a constant solar activity with drivers set to 120. The time inputs represent the fall equinox (doy = 264) at 0 hours UT.

Table 2: 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

The models are run with the aforementioned inputs held constant and each of the time-series geomagnetic indices being increased in magnitude independently. The three models using *ap* increase through the 28 discrete values between 0 and 400. For CHAMP-ML, *SYM-H* decreases from 0 to -250 in increments of 5. We then compute the density ratios for each time series index with respect to the density when it is set to 0. We plot all the curves generated for each location/model.

3 Results and Discussion

We first show the error statistics for the three ML models developed in Table 3. These are computed with respect to the original datasets. The three sets are explained in Appendix A. For reference, 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 3: Mean absolute error on the training, validation, and test sets: 100% abs(true-predicted) / true.

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%

Table 3 shows that JB2008-ML undoubtedly has the lowest errors, but it is worth noting that it is also the most generalized dataset of the three. The SET HASDM density database contains evidence of more complicated processes and its PCA coefficients are more difficult to model as a result [31]. The CHAMP-ML errors are 1–2% higher than those of HASDM-ML. CHAMP-ML also has to learn the functional relationship between density and location, which is not explicitly the case for the other two models. To both visualize the model performance in an operational setting and motivate the remainder of the work, we show the orbit-averaged densities for the three ML models and NRLMSIS 2.0 compared to the Mehta et al. (2017) CHAMP densities for the 2003 Halloween storms in Figure 1.

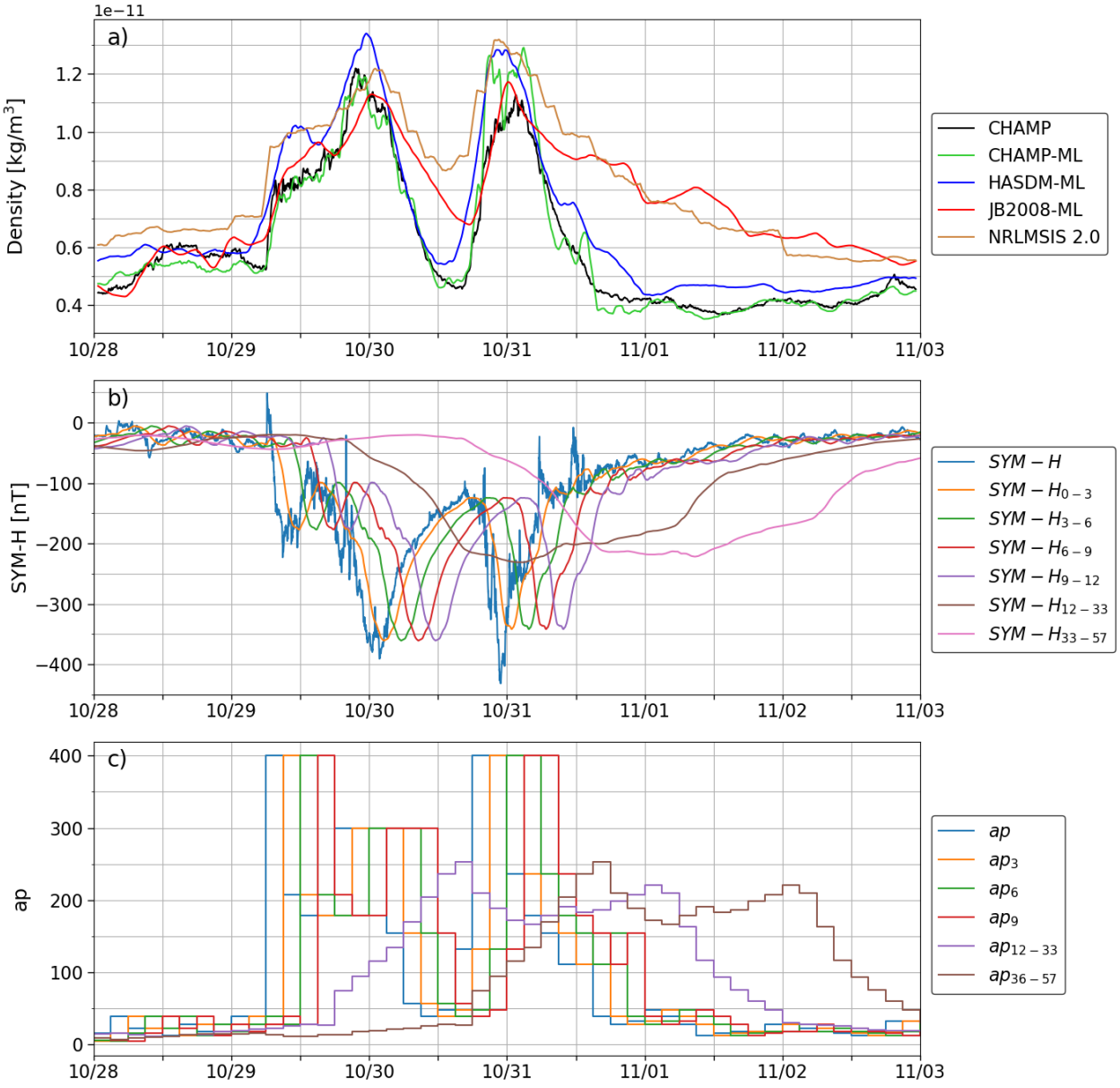

Figure 1: Orbit-average density for NRLMSIS 2.0, JB2008-ML, HASDM-ML, CHAMP-ML, and CHAMP (a) and the associated *SYM-H* (b) and *ap* (c) time-series inputs.

Figure 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). 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.

Figure 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 reinforces observations of Zesta and Oliveira (2019). 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 2.4) are displayed in Figure 2.

Figure 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.

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 $ap > 100$. While there are values for JB2008-ML in Figure 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 2. At each of the four locations, $ap_{12\text{-}33}$ and $ap_{36\text{-}57}$ have a 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.

4 Summary

In this work, we demonstrate the use of machine-learned models for investigation with a focus on thermospheric post-storm cooling. We train ML models on three datasets: JB2008 outputs, the SET HASDM density database, and the Mehta et al. (2017) CHAMP density estimates to conduct this assessment and use NRLMSIS 2.0 for comparison. All models developed are 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 (Figure 1), we show 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 and CHAMP-ML both 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 isolate the internal model formulation only as it pertains to recent magnetic perturbations. To achieve this, we hold all model drivers constant and only vary a single geomagnetic driver at a time: *ap* for NRLMSIS 2.0, JB2008-ML and HASDM-ML and *SYM-H* for CHAMP-ML. 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. In fact, the most important historical driver to JB2008-ML was the 9-hour prior *ap* which resulted in density ratios nearly twice that of any other driver.

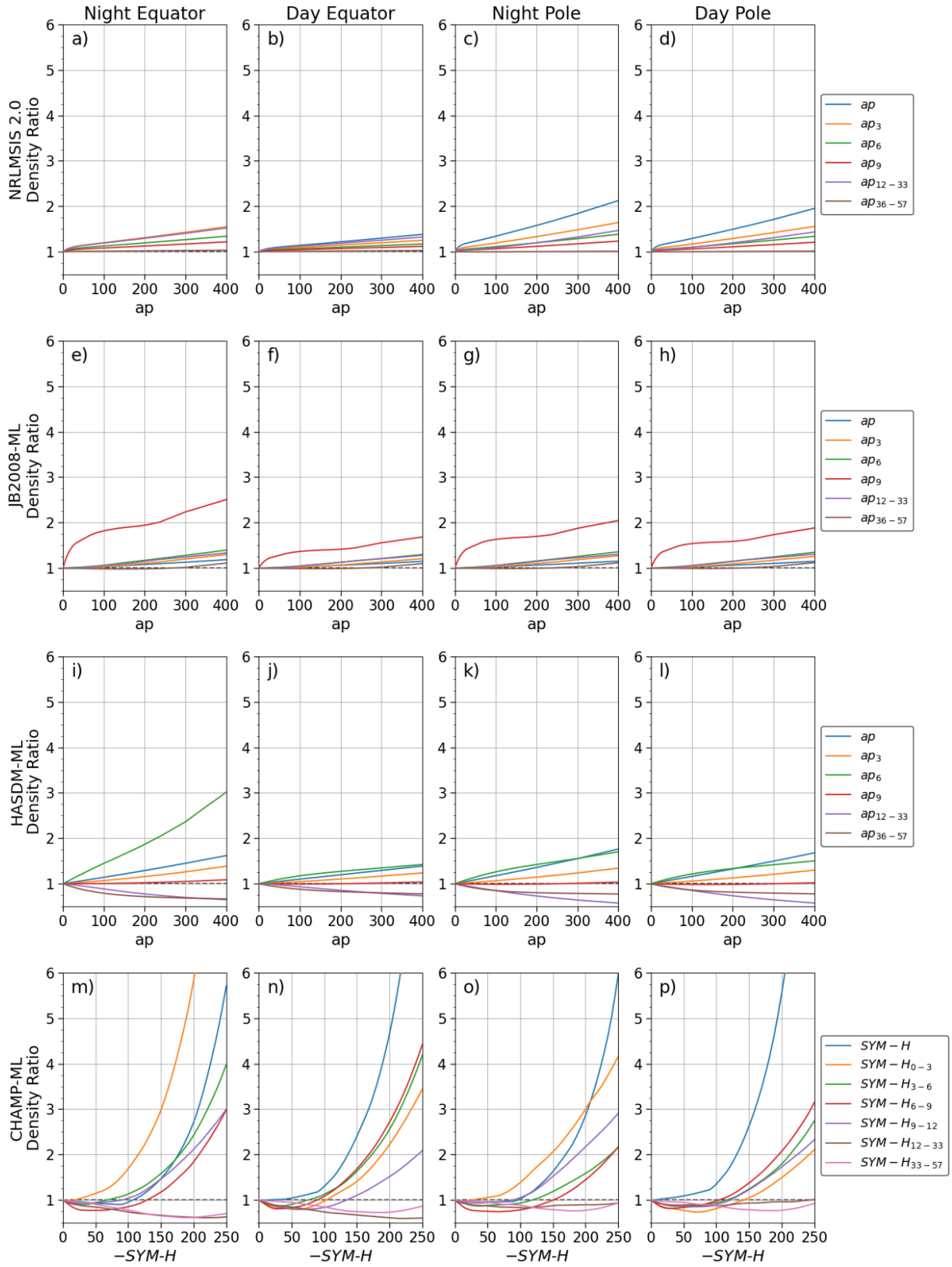

Figure 2: Density ratios for the four locations and four models, as described in Section 2.4.

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 was the only model to indicate a 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 less than 1.00. As the index was made more negative, the least recent drivers caused the lowest density ratios, particularly at low latitudes.

While the way the historical geomagnetic indices were varied may be nonphysical, this evaluates what relationship the ML models learned from their respective datasets. ML regression models are difficult to examine as it pertains to the "hidden" internal structure, but an analysis like this can inform us what the overall formulation the model has with respect to its drivers. This method can be used to gain a better understanding of the model and to investigate complex space weather datasets.

Data Statement

Requests can be submitted for full access to the SET HASDM density database at `https://spacewx.com/hasdm/` and all reasonable requests for scientific research are accepted as explained in the rules of road document on the website. The historical space weather indices used in this study can be found at the following links: $F_{10.7}$: `https://www.spaceweather.gc.ca/forecast-prevision/solar-solaire/solarflux/sx-en.php`, *ap*: `https://doi.org/10.5880/Kp.0001`, *Dst*: `http://wdc.kugi.kyoto-u.ac.jp/dstdir/`, and *SYM-H*: `http://wdc.kugi.kyoto-u.ac.jp/aeasy/index.html`. The remaining solar indices and proxies can be found at `https://spacewx.com/jb2008/` in the SOLFSMY.TXT file. Free and one-time only registration is required to access the nowcasts and forecasts. CHAMP and GRACE position data were obtained from the measurements presented by Mehta et al. (2017) at `http://tinyurl.com/densitysets`.

Acknowledgements

This work was supported by NASA grant 80NSSC20K1362 to Virginia Tech under the Space Weather Operations 2 Research Program, with subcontracts to WVU and SET. PMM gratefully acknowledges support under NSF CAREER award #2140204. SET and WVU gratefully acknowledge support from the NASA SBIR contract #80NSSC20C0292 for Machine learning Enabled Thermosphere Advanced by HASDM (META-HASDM). We would like to thank Space Weather Canada for providing and maintaining solar radio emission data, GFZ Potsdam for supplying *ap* archives, and the World Data Center for Geomagnetism in Kyoto for providing *Dst* and *SYM-H* data. The authors would like to acknowledge DLR for their work on the CHAMP mission along with GFZ Potsdam for managing the data.

References

[1] K. Chellapilla, S. Puri, and P. Simard, "High Performance Convolutional Neural Networks for Document Processing," in *Tenth International Workshop on Frontiers in Handwriting Recognition* (G. Lorette, ed.), (La Baule (France)), Université de Rennes 1, Suvisoft, 2006.

[2] K. Florios, I. Kontogiannis, S.-H. Park, J. A. Guerra, F. Benvenuto, D. S. Bloomfield, and M. K. Georgoulis, "Forecasting solar flares using magnetogram-based predictors and machine learning," *Solar Physics*, vol. 293, no. 2, pp. 1–42, 2018.

[3] Y. Jiao, J. J. Hall, and Y. T. Morton, ""automatic equatorial gps amplitude scintillation detection using a machine learning algorithm"," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 1, pp. 405–418, 2017.

[4] S. Wing, J. R. Johnson, J. Jen, C.-I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello, M. Balikhin, and K. Takahashi, "Kp forecast models," *Journal of Geophysical Research: Space Physics*, vol. 110, no. A4, 2005.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ""imagenet classification with deep convolutional neural networks"," in *"Advances in Neural Information Processing Systems"* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[6] J. Emmert, "Thermospheric mass density: A review," *Advances in Space Research*, vol. 56, 06 2015.

[7] T. J. Fuller-Rowell, M. V. Codrescu, R. G. Roble, and A. D. Richmond, *How Does the Thermosphere and Ionosphere React to a Geomagnetic Storm?*, pp. 203–225. American Geophysical Union (AGU), 1997.

[8] E. Zesta and D. M. Oliveira, "Thermospheric heating and cooling times during geomagnetic storms, including extreme events," *Geophysical Research Letters*, vol. 46, no. 22, pp. 12739–12746, 2019.

[9] G. Kockarts, "Nitric oxide cooling in the terrestrial thermosphere," *Geophysical Research Letters*, vol. 7, no. 2, pp. 137–140, 1980.

[10] J. M. I. Russell, M. G. Mlynczak, L. L. Gordley, J. Tansock, and R. Esplin, "An Overview of the SABER Experiment and Preliminary Calibration Results," *Space Dynamics Lab Publications*, vol. 114, 1999.

[11] C. Reigber, H. Lühr, and P. Schwintzer, "Champ mission status," *Advances in space research*, vol. 30, no. 2, pp. 129–134, 2002.

[12] S. Bettadpur, "Gravity Recovery and Climate Experiment: Product Specification Document," *GRACE 327-720, CSR-GR-03-02*, 2012. Cent. for Space Res., The Univ. of Texas, Austin, TX, https://podaac.jpl.nasa.gov/GRACE.

[13] M. Mlynczak, F. J. Martin-Torres, J. Russell, K. Beaumont, S. Jacobson, J. Kozyra, M. Lopez-Puertas, B. Funke, C. Mertens, L. Gordley, R. Picard, J. Winick, P. Wintersteiner, and L. Paxton, "The natural thermostat of nitric oxide emission at 5.3 μm in the thermosphere observed during the solar storms of April 2002," *Geophysical Research Letters*, vol. 30, no. 21, 2003.

[14] J. Lei, A. G. Burns, J. P. Thayer, W. Wang, M. G. Mlynczak, L. A. Hunt, X. Dou, and E. Sutton, "Overcooling in the upper thermosphere during the recovery phase of the 2003 October storms," *Journal of Geophysical Research: Space Physics*, vol. 117, no. A3, 2012.

[15] D. J. Knipp, D. V. Pette, L. M. Kilcommons, T. L. Isaacs, A. A. Cruz, M. G. Mlynczak, L. A. Hunt, and C. Y. Lin, "Thermospheric nitric oxide response to shock-led storms," *Space Weather*, vol. 15, no. 2, pp. 325–342, 2017.

[16] A. V. Mikhailov and L. Perrone, "Poststorm Thermospheric NO Overcooling?," *Journal of Geophysical Research: Space Physics*, vol. 125, no. 1, p. e2019JA027122, 2020.

[17] J. Lei, W. Wang, A. G. Burns, S.-R. Zhang, and T. Dang, "Comments on "Poststorm Thermospheric NO Overcooling?" by Mikhailov and Perrone (2020)," *Journal of Geophysical Research: Space Physics*, vol. 126, no. 4, p. e2020JA027992, 2021.

[18] J. T. Emmert, D. P. Drob, J. M. Picone, D. E. Siskind, M. Jones Jr., M. G. Mlynczak, P. F. Bernath, X. Chu, E. Doornbos, B. Funke, L. P. Goncharenko, M. E. Hervig, M. J. Schwartz, P. E. Sheese, F. Vargas, B. P. Williams, and T. Yuan, "NRLMSIS 2.0: A Whole-Atmosphere Empirical Model of Temperature and Neutral Species Densities," *Earth and Space Science*, vol. 8, no. 3, p. e2020EA001321, 2021. e2020EA001321 2020EA001321.

[19] A. E. Hedin, "MSIS-86 Thermospheric Model," *Journal of Geophysical Research: Space Physics*, vol. 92, no. A5, pp. 4649–4662, 1987.

[20] B. Bowman, W. K. Tobiska, F. Marcos, C. Huang, C. Lin, and W. Burke, "A New Empirical Thermospheric Density Model JB2008 Using New Solar and Geomagnetic Indices," in *AIAA/AAS Astrodynamics Specialist Conference*, AIAA 2008-6438, 2008. https://arc.aiaa.org/doi/abs/10.2514/6.2008-6438.

[21] A. E. Covington, "Solar Noise Observations on 10.7 Centimeters," *Proceedings of the IRE*, vol. 36, no. 4, pp. 454–457, 1948.

[22] W. K. Tobiska, S. D. Bouwer, and B. R. Bowman, "The development of new solar indices for use in thermospheric density modeling," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 70, no. 5, pp. 803–819, 2008.

[23] W. K. Tobiska, B. R. Bowman, D. Bouwer, A. Cruz, K. Wahl, M. Pilinski, P. M. Mehta, and R. J. Licata, "The SET HASDM density database," *Space Weather*, p. e2020SW002682, 2021.

[24] M. F. Storz, B. R. Bowman, M. J. I. Branson, S. J. Casali, and W. K. Tobiska, "High accuracy satellite drag model (hasdm)," *Advances in Space Research*, vol. 36, no. 12, pp. 2497–2505, 2005.

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

[26] S. Bruinsma and R. Biancale, "Total Densities Derived from Accelerometer Data," *Journal of Spacecraft and Rockets*, vol. 40, no. 2, pp. 230–236, 2003.

[27] H. Liu, H. Lühr, V. Henize, and W. Köhler, "Global distribution of the thermospheric total mass density derived from CHAMP," *Journal of Geophysical Research: Space Physics*, vol. 110, no. A4, 2005.

[28] E. 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. https://ui.adsabs.harvard.edu/abs/2008PhDT.....87S.

[29] E. Doornbos, *Producing Density and Crosswind Data from Satellite Dynamics Observations*, pp. 91–126. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[30] G. March, E. Doornbos, and P. Visser, "High-fidelity geometry models for improving the consistency of CHAMP, GRACE, GOCE and Swarm thermospheric density data sets," *Advances in Space Research*, vol. 63, no. 1, pp. 213–238, 2019.

[31] R. J. Licata, P. M. Mehta, W. K. Tobiska, B. R. Bowman, and M. D. Pilinski, "Qualitative and Quantitative Assessment of the SET HASDM Database," *Space Weather*, vol. 19, no. 8, p. e2021SW002798, 2021.

[32] P. M. Mehta and R. Linares, "A methodology for reduced order modeling and calibration of the upper atmosphere," *Space Weather*, vol. 15, no. 10, pp. 1270–1287, 2017.

[33] P. M. Mehta, R. Linares, and E. K. Sutton, "A Quasi-Physical Dynamic Reduced Order Model for Thermospheric Mass Density via Hermitian Space-Dynamic Mode Decomposition," *Space Weather*, vol. 16, no. 5, pp. 569–588, 2018.

[34] R. J. Licata, P. M. Mehta, W. K. Tobiska, and S. Huzurbazar, "Machine-Learned HASDM Model with Uncertainty Quantification," 2021.

[35] T. Iyemori, "Storm-Time Magnetospheric Currents Inferred from Mid-Latitude Geomagnetic Field Variations," *Journal of geomagnetism and geoelectricity*, vol. 42, no. 11, pp. 1249–1265, 1990.

[36] F. Siciliano, G. Consolini, R. Tozzi, M. Gentili, F. Giannattasio, and P. De Michelis, "Forecasting SYM-H Index: A Comparison Between Long Short-Term Memory and Convolutional Neural Networks," *Space Weather*, vol. 19, no. 2, p. e2020SW002589, 2021.

[37] D. R. Weimer, "Improved ionospheric electrodynamic models and application to calculating Joule heating rates," *Journal of Geophysical Research: Space Physics*, vol. 110, no. A5, 2005.

[38] D. R. Weimer, "Predicting surface geomagnetic variations using ionospheric electrodynamic models," *Journal of Geophysical Research: Space Physics*, vol. 110, no. A12, 2005.

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

[40] R. J. Licata and P. M. Mehta, "Uncertainty Quantification Techniques for Space Weather Modeling: Thermospheric Density Application," 2022.

Appendix A: Data Splitting

When developing a ML model, one must carefully consider how to split the data into training, validation, and test sets. These models can have thousands or even millions of parameters, so it is pertinent to ensure the analysis is valid and that the model has not memorized the training data. The goal is to provide enough diverse data in the training set for the model to learn the proper relationship between the inputs and outputs. The validation and test sets are chosen to also contain diverse condition. The validation set is used to determine the model performance – during and after training – on data not used for fitting. However, since you choose the best version of the model based on its training and validation performance, there must be another independent set for model evaluation; this is called the test set.

For JB2008-ML and HASDM-ML, the data splitting is straightforward. Looking at the solar and geomagnetic drivers, we split the data by years trying to place various portions of the solar cycle in each set. The breakdown is displayed in Table A1. This results in 12 years being used for training (60%), 4 years being used for validation (20%), and 4 years being used for test (20%).

Table A1: Training, validation, and test split for JB2008-ML and HASDM-ML by year.

Set	Training	Validation	Test
Years	2000, 2001, 2003, 2004, 2007, 2009, 2011, 2012, 2013, 2015, 2016, 2019	2005, 2008, 2014, 2017	2002, 2006, 2010, 2018

For CHAMP-ML, a combination of having location as an input and not having a full solar cycle of data makes the data splitting more complex. Using year-long periods (as in Table A1) leaves out an entire range of possible altitudes and solar drivers from training. We work around this by repeating the following scheme: 8 consecutive weeks for training, 1 week for validation, and 1 week for test. Even though there are only two weeks between training segments, the 10-second cadence of the measurements creates a gap of over 120,000 samples, ensuring the model is not just simply interpolating for the validation and test sets. While only 60% of data is in the JB2008-ML and HASDM-ML training sets, the spatiotemporally limited CHAMP dataset requires more training data to prevent overfitting.

Appendix B: Additional Model Development Details

Keras Tuner

As discussed in Section 2.2, we use Keras Tuner to identify an architecture for each dataset. The hyperparameter space we define for the tuners (one for each dataset) is displayed in Table B1.

Table B1: Hyperparameter search space for the three ML models.

Parameter	Choices
Number of Hidden Layers	1 – 10
Neurons	min = 64, max = 1024, step = 4
Activations	relu, softplus, tanh, sigmoid, softsign, selu, elu, linear
Dropout	min = 0.10, max = 0.60, step = 0.01
Optimizer	RMSprop, Adam, Nadam, Adadelta, Adagrad

For JB2008-ML and HASDM-ML, the tuner is provided the entire training and validation sets due to the relatively small size of the datasets. Therefore, once the tuner returns the best models, each trained for 100 training iterations or epochs, there is no need for further training. In fact, training after tuning did not yield improved models in these two cases. The CHAMP-ML training and validation sets have over 20 million and 2 million samples, respectively. Therefore, we only supply the tuner with 1 million random samples from the training set and 200,000 random samples from the validation set. This provides the tuner with ample information to determine an adequate architecture from which we can develop the full model. This full model will use the full training and validation sets.

Loss Function and Custom Layer

The loss function plays an important role in the tuning and training process. It is the metric that the fitting function tries to minimize or maximize through the weight updates. For this study, a mean square error (MSE) loss function would suffice since the goal is to develop a model with minimal error with respect to the respective dataset. However, we used models from previous work, Licata et al (2022), which used the negative logarithm of predictive density (NLPD) loss function,

$$NLPD(y, \mu, \sigma) = \frac{(y-\mu)^2}{2\sigma^2} + \frac{ln(\sigma^2)}{2} + \frac{ln(2\pi)}{2} \tag{3}$$

where y, μ, and σ are the ground truth, mean prediction, and standard deviation of the prediction, respectively. In practice, the factor of 1/2 and the last term can be removed. There are no ground truth values for σ for these datasets, but we are able to directly predict it since it is a standalone term in Equation 3. This is accomplished by using twice the output neurons as there are outputs where they represent the mean and standard deviation of the model uncertainty. The output neurons representing the mean use a linear activation function, while those representing the standard deviation use the "softplus" activation function. The softplus activation and its derivative, the sigmoid function, are defined as,

$$f(x) = ln(1 + e^x) \qquad f'(x) = \frac{e^x}{1 + e^x} \tag{4}$$

with x being the weighted sum of the neurons inputs. In Licata et al. (2021), they found that for the SET HASDM density database, the use of NLPD as opposed to MSE did not result in significantly different prediction errors [34].

Batch Size for Tuners and Further Training

Batch size is the number of samples used in fitting to average the loss over. With too small of a batch size, the model can update its weights often to specific batches of the data making it unstable. With too large of a batch size, the model can have trouble learning due to the over-generalization of the losses. Choosing a batch size for a given application can require trial and error as there is no universal choice. Table C1 shows the batch sizes used for the tuners and subsequent training for each dataset.

Table C1: Batch sizes used to tune and train the three ML models in this work.

	JB2008-ML / HASDM-ML	CHAMP-ML
Tuner	2^8 = 256	2^{12} = 4,096
Further Training	N/A	2^{17} = 131,072

As discussed, the choices in Table C1 were chosen from trial and error. The JB2008-ML and HASDM-ML batch size is the smallest partially due to the size of the dataset (35,064 training samples). For the CHAMP-ML tuner, we use 1 million samples and therefore have a larger batch size. The batch size for training the final CHAMP-ML model using the best architecture from the tuner is significantly larger, because we have over 20 million training samples and found this leads to the most stable training process.