Quantifying the storm-time thermospheric neutral density variations using model and observations

E. Ceren Kalafatoglu Eyiguler[1], J. S. Shim[2], M. Kuznetsova[3], Z. Kaymaz[1], B. R. Bowman[4], M. V. Codrescu[5], S. C. Solomon[6], T. J. Fuller-Rowell[5], A. J. Ridley[7], P. M. Mehta[8] and E. K. Sutton[9]

[1] İstanbul Technical University, Faculty of Aeronautics and Astronautics, Department of Meteorology, İstanbul, TR

[2] The Catholic University of America, NASA GSFC, Greenbelt, MD, USA

[3] NASA, Goddard Space Flight Center, Greenbelt, MD, USA.

[4] Air Force Space Command Space Analysis Division, Peterson AFB, CO, USA

[5] Space Weather Prediction Center, NOAA, Boulder, CO, USA

[6] High Altitude Observatory, National Center for Atmospheric Research, Boulder, CO, USA

[7] School of Engineering, University of Michigan, Ann Arbor, MI, USA

[8] Department of Mechanical and Aerospace Engineering, West Virginia University, WV, USA.

[9] Air Force Research Laboratory, Kirtland AFB, NM, United States

Corresponding author: E. Ceren Kalafatoglu Eyiguler (ceren.kalafatoglu@itu.edu.tr)

Key Points:

- Using the average and maximum values of neutral densities to determine the model performances can be misleading
- Removing the quiet-time trend from the neutral density reveals the actual performance of the model in simulating the storm-time variations
- Mean absolute error, prediction efficiency and normalized root mean square error should be considered together for the evaluations

Abstract

Accurate determination of thermospheric neutral density holds crucial importance for satellite drag calculations. The problem is two-fold and involves the correct estimation of the quiet-time climatology and storm-time variations. In this work, neutral density estimations from two empirical and three physics-based models of the ionosphere-thermosphere are compared with the neutral densities along the CHAMP satellite track for six geomagnetic storms. Storm time variations are extracted from neutral density by 1) subtracting the mean difference between model and observation (bias), 2) setting climatological variations to zero, and 3) multiplying model data with the quiet time ratio between the model and observation. Several metrics are employed to evaluate the model performances. We find that the removal of bias or climatology reveals actual performance of the model in simulating the storm-time variations. When bias is removed, depending on event and model, storm-time errors in neutral density can decrease by an amount of 113% or can increase by an amount of 12% with respect to error in models with quiet time bias. It is shown that using only average and maximum values of neutral density to determine the model performances can be misleading since a model can estimate the averages fairly-well but may not capture the maximum value or vice versa. Since each of the metrics used for determining model performances provide different aspects of the error, among these, we suggest employing mean absolute error, prediction efficiency and normalized root mean square error together as standard set of metrics for the neutral density.

Plain Language Summary

Thermospheric neutral density is the largest source of uncertainty in atmospheric drag calculations. Consequently, mission and maneuver planning, satellite lifetime predictions, collision avoidance and orbit determination depend on the accurate estimation of the thermospheric neutral density. Thermospheric neutral density varies in different time scales. In short time scales, the largest variations occur due to the geomagnetic storms. Several empirical and physics-based models of the ionosphere-thermosphere system are used for estimating the variations in the neutral density. However, the storm-time responses from the models are clouded by the climatology (background variations), upon which the effect of geomagnetic storms are superimposed. In this work, we show that it is critical to use reference levels for the neutral density to extract the true performance of the models for the evaluation of the storm-time performances. We demonstrate that mean absolute error, prediction efficiency and normalized root mean square error should be considered together for the performance evaluations, since each of them provides different aspects of the error.

1 Introduction

It is known that the atmospheric drag acting on satellites is significant between the altitudes 160 and 800 km (Zesta and Huang, 2016). Consequently, in atmospheric drag calculations, in orbit determination, the largest uncertainty comes from the thermospheric neutral density (Hejduk and Snow, 2018; Bussy-Virat et al., 2017). The effects of the uncertainty in neutral density are not only limited to orbit prediction, accurate density estimates are also needed for mission and maneuver planning and collision avoidance (Storz et al., 2005). Low Earth orbit (LEO) satellites are under the influence of the thermospheric environment and their lifetimes depend on the variation of the neutral density (Prölss, 2011). Consequently, real-time estimation of the atmospheric drag, which is important for satellite operations, heavily relies on the correct estimation of the thermospheric neutral density.

Variations in thermospheric density can be decomposed into three main components: 1) the variations, which are governed by the solar irradiance (solar-cycle dependent, seasonal, diurnal)

(Qian and Solomon, 2012), 2) the variations due to upward propagating tides and waves from the mesosphere (Sutton et al., 2005), and 3) the storm-time variations, which are largely influenced by the heat sources that come into play during geomagnetic activity, such as Joule heating (Fedrizzi et al, 2011; Kim et al., 2006), auroral particle precipitation (Deng et al., 2013) and heating due to small scale field-aligned currents (FACs) (Lühr et al., 2004). The former two components control the quiet-time variation in neutral density, which is referred to as climatological (background) variations in this study. In addition, the thermospheric composition modulates the changes in thermospheric neutral density (Qian et al., 2009). In some geomagnetic storm cases, the damping of the thermospheric density by NO cooling is significantly stronger than expected. Those cases are classified as problem storms by Knipp et al. (2013) and it is shown that the thermosphere's response is strongly associated with the pre-storm properties of the solar wind. Different drivers of geomagnetic storms, such as the Coronal Mass Ejections (CME) and Corotating Interaction Regions (CIRs) cause different environmental responses in the thermosphere (McGranaghan et al., 2014). CIR and CME effects on thermospheric densities were investigated in several studies (Chen et al., 2014; Chen et al., 2012; McGranaghan et al., 2014; Lei et al., 2011; Thayer et al., 2008). Even though less geoeffective in terms of Dst magnitude, the total effect of CIR storms was found to be comparable to CME induced enhancements in thermospheric neutral density (Chen et al., 2014).

LEO satellite observations and empirical and physics-based models are employed in the investigations of thermospheric neutral density (Lathuillère et al., 2008; Sutton et al., 2006; Liu et al., 2005; Pardini et al., 2012; Codrescu et al., 2012; Deng et al., 2013; Solomon et al, 2011). The Challenging Micro-Satellite Payload (CHAMP) and Gravity Recovery and Climate Experiment (GRACE) satellites are the most used satellites for the investigations of the neutral density and the associated atmospheric drag acting on satellites (Anderson et al., 2009; Picone et al., 2002; Bruinsma and Forbes, 2010; Liu et al., 2010; Xu et al., 2011; Huang et al., 2014; Bruinsma, 2015; Bruinsma et al., 2018). Recently, data from Swarm constellation has also been employed to derive the thermospheric neutral densities (Huang and Zesta, 2016; Siemes et al., 2016; Kodikara et al., 2018). In this kind of approach, the densities are calculated from the accelerometers on the spacecraft (Sutton, 2005).

However, in-situ measurements from satellites only provide the current state of the thermosphere. Hence, the empirical models involving semi-physical relations, which take geomagnetic and solar indices as input and the physics-based models of the ionosphere-thermosphere (IT) are employed to nowcast and forecast of the future state of the IT system in global scales. The nowcast and forecast of neutral density are necessities for early-action and response and orbit determination of the LEO spacecraft.

Comparisons between the model and observations are made in different time scales: daily global mean (Solomon et al., 2011; Qian et al., 2008), orbit averaged (Bowman et al., 2008) and along the satellite track (Connor et al., 2016; Shim et al., 2012). Comparisons for longer time scales that are associated with the periodicities in neutral density such as the 27-day, 81-day and yearly variations were also carried out in several studies (Rhoden et al., 2000; Qian and Solomon, 2012; Bruinsma et al., 2018).

Several metrics are employed to assess the model performances. For the neutral density studies, the most used metrics are the mean absolute error (MAE), bias (B), correlation (R), root mean square error (RMSE), standard deviation (Std), prediction efficiency (PE), ratio of maximum and ratio of average (Pardini et al., 2012, Shim et al., 2012; Elvidge et al., 2014; Bruinsma, 2015; Elvidge et al., 2016; Emmert et al., 2017; Kodikara et al., 2018) and the version of the metrics in log space (Picone et al., 2002; Sutton, 2018; Bruinsma et al., 2018). Each of these metrics has advantages and disadvantages (Hyndman et al., 2012; Shcherbakov et al., 2013). For example, the MAE provides the average difference between the model and observation and it is easy to use. However, it does not offer any information on the amount of the error when compared to the variations at large with respect to the event in percentage. Likewise, "ratio"s provide the difference between the observation and estimate at an instant, but they do not deliver information on the properties of the temporal evolution of the error. Std and RMSE are highly sensitive to outliers and may lead to the overestimation of errors in some cases. Among the metrics, the PE is becoming increasingly used by the space weather community. PE is a dimensionless quantity and represents the measure of success in reproducing a time series. PE basically compares the order of magnitude of model errors with the magnitude of variations of the measurements/reference data. However, one handicap of PE is that, it does not provide the actual value of difference between the observation and estimations. It is also worth to note that in the literature, same equations are used in the calculations of all metrics given above, except the bias metric. Bias may have different definitions based on the study. Bias is sometimes calculated as the difference between the model and observation in percentage (Pardini et al., 2012) and sometimes as the mean difference between the model and observation (Elvidge et al., 2016). In our work, we define model bias as the quiet-time mean difference between the model and observation (mean of model minus mean of observation). Additionally, we do not use it as a metric, but rather, use the quiet-time model bias to extract the storm-time variations from the neutral density. The definitions of the metrics that we use in our study are given in Section 2.3.

As a summary, all metrics provide different aspects of the error. Hence, Chai et al. (2014) suggests using not only one, but several metrics together, especially in studies involving the assessment of more than one model when the error distribution becomes important. Consequently, this is the case for the neutral density studies and a variety of metrics are employed together in comparisons. However, there are not any consensus on what to use as a standard set of metrics. The community need at the current time is to be able to run the models for real-time calculations of atmospheric drag in support of real-time satellite operations. For this purpose, there is a need to assess the performances of the models and to specify the conditions when they perform satisfactorily and when they do not (Shim et al., 2014; Shim et al., 2015). This study is a continuation of the GEM-CEDAR challenge for the assessment and benchmarking of the empirical and coupled models of the ionosphere-thermosphere and is a deliverable of the International Forum on Space Weather Modeling Capabilities Assessment. In the first study of the series, Shim et al. (2011) compared the model results with the local measurements available from EISCAT radars for the ionospheric parameters N_mF2, h_mF2 and vertical drift with limited latitudinal coverage. Shim et al. (2012) focused on the space-borne measurements of the N_mF2, h_mF2, ionospheric electron density and thermospheric neutral density along the satellite track at the measurement locations. N_mF2 and h_mF2 from the models were compared with the observations from the Constellation Observing System for Meteorology, Ionosphere and Climate (COSMIC) while ionospheric electron density and thermospheric neutral density were compared using the measurements from CHAMP. In both studies, root mean square error (RMSE), prediction efficiency (PE), ratio of (max-min) and ratio of maxima were employed to assess the model performances. They reported that the model performances depend on the metrics used and varied with latitude and geomagnetic levels. No models outperformed others in estimating the thermospheric and ionospheric parameters in all cases.

In model comparison and validation, the absence of a standard set of metrics complicates the evaluation and synthesis of the results of different studies. As a part of the systematic evaluation of the models in this study, our aims are to present ways to facilitate the comparison of the storm-time performances of the models and to provide a useful set of metrics for the neutral density studies. We present methods to remove the quiet time variations from the neutral density, so that the storm-time changes are revealed. Accordingly, direct comparisons can be made between the model estimations and observations from the CHAMP satellite for the disturbed periods. The climatology removal methods are called as baseline shifts, since they match the level of quiet-time neutral density estimated from the models with the quiet-time level of neutral density variations observed by CHAMP. Orbital averages of thermospheric neutral density along the CHAMP satellite track are used to evaluate the model performance. We show that baseline shifts are a necessity in order to correctly assess the storm-time performances of the models and the climatology and storm-time variations should be evaluated separately as the dominant mechanisms and their time-scales are different in each. In Section 2, the events selected for the case studies are introduced and baseline shifting methods are described. Section 3 presents the results and involves the comparison of baseline shifting methods and the neutral density estimations from the empirical and physics-based models of the IT. Lastly, we conclude the study and discuss the future needs of the community in Section 4.

2 Data and Methodology

Two empirical and three physics-based models are employed in this study. The empirical models are Naval Research Laboratory Mass Spectrometer and Incoherent Scatter Extended (NRLMSISE-00, will be referred to as MSIS, hereafter) (Picone et al., 2002) and Jacchia-Bowman-2008 (JB2008) (Bowman et al., 2008), whereas the physics-based models are

Thermosphere-Ionosphere-Electrodynamics General Circulation Model (TIEGCM1.95) (Richmond et al., 1992), Coupled Thermosphere Ionosphere Plasmasphere electrodynamics (CTIPe) (Millward et al., 2001; Codrescu et al., 2008) and GITM (Ridley et al., 2006). The models were run using the NASA Community Coordinated Modeling Center Runs-on-Request system. The results can be found by searching the simulation IDs that are given in Table S1. Additionally, Table S1 provides information on the version and the resolution of the models for each run. For each run and model, the initial parameters and model input are the same. Table S2 shows the input parameters to the models. For physics-based models, ionospheric electric potentials have to be specified to describe the interaction of the solar wind and magnetosphere with the ionosphere. This is handled by selecting a high-latitude driver, which describes the electrodynamic input from the magnetosphere and solar wind into the high-latitude ionosphere under different solar wind conditions. In this study, Weimer-2005 (Weimer, 2005) ionospheric potentials are employed as the high-latitude driver for each physics-based model for consistency. Details on the models and their standard configurations for the runs can be found in (Shim et al., 2011; Shim et al., 2012).

The model results are compared against the newly updated thermospheric neutral density data set from CHAMP by Mehta et al. (2017), which is referred to as M2017, hereafter. Previous studies of systematic assessment (Shim et al., 2012; Shim et al., 2014) used older versions of neutral density data that were also derived from CHAMP accelerometer measurements (Sutton et al., 2005). Besides, prior to the M2017, the most recent version of neutral density data which had been widely used in comparisons was the Version 2.3 of Sutton (2009). This version is also detailed on a report by (Sutton, 2011). The differences between the previous versions of neutral density data sets and the M2017 are associated with the modeling of the drag coefficient (C_D), which is a coefficient in the equation of satellite drag. The drag coefficient is a number that depends on the geometry of the spacecraft and the properties of the impinging particles. Precise calculations of the drag coefficient are necessary for accurate neutral density estimations, since the neutral density is calculated using accelerometer data, hence the C_D. The M2017 considers a more complicated geometry and uses the most recent advances in the modeling of gas-surface interactions and the modeling of physical C_D. In their work, Mehta et al. (2017) reported differences up to 20% for some cases with respect to the neutral density estimates of Sutton (2008). In this study, to give the difference between the newly derived and old data sets, the Version 2.3 data set (Sutton, 2009) is also included in the comparisons. The (Sutton, 2009) Version 2.3 is represented as SV2.3 throughout the paper.

In this work, we investigate the storm-time performances of the IT models for six geomagnetic storms, which were particularly chosen by the GEM-CEDAR community for the systematic evaluation of the models. According to the NOAA classification based on the Kp index, the intensity of selected events ranges from weak to severe. Table 1 presents the extreme values of geomagnetic and solar indices along with the solar wind drivers for the events Hemispheric Power (HP) index is also given in Table 1 since it is an input to the physics-based models. In the Table, HSS denotes the high speed streams.

Figure 1 shows the storm-time maximum neutral density on the left, storm-time average neutral density from the models and M2017 in the middle, and the timing difference between the neutral density maximum in M2017 and the maximum in models in the right panel, for each geomagnetic storm case. As evident from the plot, the storm-time maximum and average neutral densities from M2017 display a decreasing trend with weaker geomagnetic storms. Even though SV2.3 always shows higher values than M2017, it follows the same trend in neutral densities.

For the neutral density maximum, all models show the same tendency as in CHAMP observations, except the 2005-243 event, which is due to an HSS. TIEGCM and JB2008 overestimate the neutral density peak in each event, whereas GITM slightly underestimates in four of the six events (2005-135, 2005-243, 2007-142 and 2007-91). MSIS neutral density maxima are higher than M2017 for events with Kp<6, but lower than M2017 for events with Kp>6, except the 2006-348 event. CTIPe estimates are slightly higher than but very close to M2017 in most of the events. Overall, CTIPe and GITM are the two models that generally show the closest neutral density maxima to M2017.

These patterns in the modeled neutral density maxima change in the average neutral densities. A model overestimating the neutral density maxima in M2017 can give a lower average than the M2017 or vice versa for the same events. For example, JB2008 and GITM for 2005-135, TIEGCM for the 2005-243 and MSIS for the 2006-348 and 2007-142 show the opposite behavior in terms of storm-time neutral density average and maximum. In the figure, it is seen that MSIS underestimates the neutral density average in all selected events except the 2007-91.

JB2008 overestimates the storm-time neutral density in four of the six events and underestimates in two events. Neither the MSIS, nor the JB2008 display the decreasing trend with weakening geomagnetic activity in average neutral density average that is illustrated in M2017 for the selected event set. Despite, TIEGCM, and GITM display the decreasing trend also for the neutral density averages, except the 2005-243 event as in neutral density maxima case. None of the models are found to be consistently closer to M2017 in terms of neutral density average.

Timing differences between the models and M2017 also change with respect to event. Interestingly, most of the models performed the best in capturing the timing of maximum in 2005-190 event, which is due to a CME during an HSS. The variations in timing differences seem to be random. The timing difference between the maxima of M2017 and the models are found to be between ±7.5 hours.

In Figure 1, the storm-time neutral density maxima and averages include not only the storm-time neutral density variations but also the climatological variations. That is, the model biases are also included in evaluations. In the following sections, we show that removing the climatology or quiet-time model bias reveals the actual performance of the models in simulating the thermospheric neutral density variations during geomagnetic activity. Our approach for assessing the storm-time model performances consists of three steps, such as orbit averaging, climatology/bias removal and assessment of the results. In the following sections, we describe the tools designed for each step. The codes were written in MATLAB and are in transition to Python language.

2.1 Orbit Averaging Tool (OAT)

The orbit averaging tool (OAT) is used for taking orbital averages of thermospheric neutral density from CHAMP and models. Comparisons along the track involve local time effects, small-scale structures, and diurnal and seasonal variations (Qian and Solomon, 2012; Liu et al., 2005; Lühr et al., 2004; Kwak et al., 2009), which make it hard to specify the reason behind the difference in model estimations and observations. On the other hand, taking orbital averages smooths out the temporal and spatial variations due to the spacecraft position on a single orbit and provides the globally averaged response to the geomagnetic storm. It was also shown previously by Burke et al. (2007) that the change in orbit-averaged densities occurs systematically whereas the local density exhibits large variations.

The OAT works with CHAMP ephemeris data. First, the beginning and end times of each orbit are determined: an orbit starts at the highest northern latitude, crosses the highest southern latitude and ends at the highest northern latitude. One orbit lasts approximately 92 minutes. There are typically ~15 orbits in a day. Neutral density observations from CHAMP and estimations from each model are averaged over every single orbit of the spacecraft.

2.2 Baseline Shifting Tool (BAST)

In this study, we are concerned with the storm-time performances of the models. Thus, to compare only the storm-time responses, the baseline shifting tool (BAST) is used. BAST adjusts the quiet-time neutral density level of the models to match the quiet time level of M2017. The adjustment is handled by assuming that unless there is a geomagnetic storm, the neutral density variations will continue to fluctuate around the quiet time level of neutral density. Consequently, three types of adjustment are employed:1) subtracting the average quiet-time difference between the models and observation (Shift1-SH1), 2) setting off the climatology to zero by subtracting the quiet-time neutral density average from the models and the observation (Shift2: SH2), and 3) multiplying the model results with the quiet time average ratio between the model and observation (Shift3:SH3). All adjustments are applied separately to the model results. Hereafter, we call the adjustments as baseline shifts, since they shift the quiet time reference level of the model results to the observation or to the zero level. In the shifting procedure, the "quiet time" refers to the neutral density variations, which are only due to the changes in the solar irradiance and tides. Subsequently, any additional changes in the neutral density that are due to the geomagnetic disturbances are referred to as storm-time variations. The storm-time variations are considered to be superimposed on the quiet time neutral density variations (Lühr et al., 2011).

All three shifts work with the quiet time average of thermospheric neutral density from the model and observations. Hence, the correct identification of the quiet time intervals is important. To determine the quiet time intervals, we select a threshold for the Kp index and the neutral density fluctuations as observed by the CHAMP satellite. An interval is defined as quiet when Kp < 3- and the orbit-averaged neutral density difference between two consecutive orbits of CHAMP is less than or equal to 1.25×10-13 kg/m^3. The threshold, 1.25×10^{-13} kg/m^3, was selected by inspecting the orbit-averaged neutral density variations on quiet day cases (2007-79, 2007-190, 2007-341) used in (Shim et al., 2012) (see Figure S1). We define it as the start of the storm when the increase in CHAMP neutral density is more than 1.25×10^{-13} kg/m^3 and there is an increasing trend in orbit-averaged neutral density in two consecutive orbits. The end of the storm is marked as the time when CHAMP neutral densities return to quiet-time average neutral density level. Table 2 details the shifts that are applied to the thermospheric neutral density.

As a result of the shifting processes, we estimate the errors to be as high as the selected threshold: $\pm$ 1.25×10^{-13} kg/m^3, which is about 5% to 7% of the quiet-time neutral density of the selected events.

Figure 2 shows the 2006-348 event, which is classified as ‘severe’ according to the NOAA geomagnetic storm scale based on Kp, as an example event for baseline shifts. The selected quiet time interval for the event, which was determined according to thresholds for Kp and neutral density level is between 13/12/2006 15:00 UT and 14/12/2006 14:00. The original time series from the model and observations are displayed on the left and the shifts 1, 2 and 3 are found on the right panels. It is seen that most of the models overestimate the neutral densities during the quiet-time interval. Appropriately, the shifts remove the bias from the models, so that we can compare the storm-time variations directly between the models and M2017.

Before the baseline shifting procedure, MSIS is one of the best performing models with a maximum close to the M2017 for the 2006-348 event. However, with the removal of its bias, it is found that it actually underestimates the neutral density enhancement due to the geomagnetic storm. In the case of TIEGCM, the model overestimates the quiet-time neutral density so much that, the neutral density maximum and average during the storm are the highest among the models. Consequently, the resulting differences between the model and observation are the highest when the quiet-time bias is included. On the other hand, shifting the baseline to M2017 levels as seen in panels b and c indicate that the storm-time response as modeled by the TIEGCM is closer to M2017 than they are before the shift. These cases demonstrate the usefulness of the shifts in determining the actual storm-time response from the models.

Following the same assumptions as in case of SH1, SH2 and SH3, several other types of shifts can also be applied to the data to remove the influence of the quiet time bias on the storm-time performances. For example, an artificial time series can be produced using the quiet-time data by assuming that the neutral density levels will remain the same on the following day. The easiest way to produce an artificial time series is to sequentially iterate the neutral density during the quiet time period to cover the entire event interval. Afterwards, this newly generated time series can be used for point-to-point subtraction of 1) bias (Shift4, SH4) and 2) quiet time neutral density at the same instant (Shift5-SH5) or for 3) point to point multiplication using the quiet time ratios (Shift6-SH6). These procedures were also investigated in this work. However, since the results of point-to-point shifts are similar to shifts based on quiet time averages, which are described above, we chose to present only the results from SH1, SH2 and SH3. However, the results of all shifts for the selected events are provided in the supplement from Figure S2 to

Figure S7. The figures demonstrate that, point-to-point shifting processes may lead to unphysical variations in neutral density as in the case of GITM for weak events in this study.

2.3 Performance Assessment Tool (PAT)

After adjusting the baseline of the model and observations, storm-time model performances are evaluated according to the M2017 data set. Performance Assessment Tool (PAT) measures the model performances during individual events according to seven metrics. Those are: ratio between the model maximum and CHAMP maximum (Ratio$_{\text{max}}$), ratio between the model mean and CHAMP mean (Ratio$_{\text{avg}}$), time delay between the peak of the model and peak of the CHAMP observation (TD), mean absolute error (MAE), normalized root mean square error (NRMSE), prediction efficiency (PE) and integrated density change (IDC). Equations from 1 to 7 show the definitions of the metrics. The subscripts "i" and "j" represent the orbit number during the quiet-time and entire event and "t", the time of the orbit, respectively. All calculations are based on the storm-time variations after performing the baseline shifts.

$$\text{Ratio}_{max} = \frac{\rho_{model,max}}{\rho_{M2017,max}} \qquad (1)$$

$$\text{Ratio}_{avg} = \frac{\rho_{model,avg}}{\rho_{M2017,avg}} \qquad (2)$$

$$\text{TD} = \; t_{model,max} - t_{M2017,max} \qquad (3)$$

$$\text{MAE} = \sum|\rho_{M2017,i} - \rho_{model,i}|\,/N \qquad (4)$$

$$\text{NRMSE} = RMSE/(\rho_{M2017,max} - \rho_{M2017,min}) = \sqrt{\sum\frac{(\rho_{M2017,i}-\rho_{model,i})^2}{N}}/(\rho_{M2017,max} - \rho_{M2017,min})$$

(5)

$$PE = 1 - RMS_{model}/RMS_{M2017} = 1 - \sqrt{\frac{\sum(\rho_{M2017,i}-\rho_{model,i})^2}{\sum(\rho_{M2017,i}-\overline{\rho}_{M2017,i})^2}} \qquad (6)$$

$$\mathrm{IDC} = \sum_{j=1}^{n_{orbit}} \Big(\sum_{t_{start}}^{t_{end}} \rho_{data,t} - \rho_{baseline}\Big)_j ; \rho_{baseline} = \sum_{i=1}^{q_{orbit}} \Big(\sum_{t_{start}}^{t_{end}} \rho_{data,t}\Big)_i / q_{orbit} \qquad (7)$$

Among the metrics, the IDC works with the orbit and storm-time integrated neutral densities. The subscript "data" in Equation 7 denotes model or M2017 data. $q_{orbit}$is the total number of orbits during the quiet time interval. n_{orbit} is the total number of orbits during the entire event and "t_{end}" and "t_{start}" denote the start and end times of each orbit. Accordingly, $\rho_{baseline}$ represents the average of the orbit-integrated neutral density during the quiet time.

In contrast, other metrics use the orbit-averaged neutral densities. The perfect score for the ratios (Ratio_{max}, Ratio_{avg}) is 1, whereas TD should be zero, meaning there is no lag between the peak of the model and the time of the maximum from CHAMP. Determining the TD for less intense events is different from determining the TD for intense events. In intense events, the maximum of the neutral density is distinguishable, whereas in less intense events, there may be numerous local maxima. Consequently, we first mark the timing of the maximum neutral density from M2017, then detect the timing of the closest local maxima from the models. MAE gives the average distance between the observation and model estimations. Values approaching to zero indicate better agreement between the model and observations. Furthermore, MAE gives a dimensioned skill score, that is, it has the same units with the neutral density (kg/m^3). On the other hand, ratios, NRMSE and PE are dimensionless. PE varies between 1 and negative infinity. PE equals to 1 indicates perfect agreement between the model and observations whereas PE=0 means the model errors are in the same order with the variations of the observations. Negative PE values show that the observed mean is a better estimate for forecasts than the model (Shim et al., 2012). The NRMSE, is the normalized version of RMSE. The NRMSE gives errors in percentage. RMSE, consequently, NRMSE, vary with the variability of error magnitudes and the mean absolute error (Wilmott and Matsuura, 2005). When interpreted together with the MAE, NRMSE provides information on the variability of error magnitudes.

3 Results and Discussion

In this section, we present the storm-time performances of the models after the baseline shifting methods are applied to the observation and model neutral density estimations from the models.

Figure 3 presents the ratio of maximum neutral density (top row) and ratio of average neutral density (middle row) from each model to M2017. The best agreements are displayed between the SV2.3 and M2017 for all events before and after the baseline shifting. The SH3 yields the best results among the shifts for the SV2.3 and lead to one-to-one match between the M2017 and SV2.3 for all events. This is because M2017 and SV2.3 are only different by a constant factor in each event and SH3 finds and removes this factor by using the ratio between the SV2.3 and M2017 during the selected quiet time interval.

For MSIS, CTIPe and GITM, baseline shifting causes the ratio of maximum to diverge from 1 for some events, whereas for TIEGCM and JB2008 the shifts cause performance enhancement in capturing the maximum in M2017. MSIS and GITM are found to underestimate the maximum in M2017 generally, after the shifts. For all models, SH1 produces the closest ratios to 1 among the shifts for both the ratio of maximum and ratio of average neutral densities. SH2 causes the ratios to be more spread for all events and models. Using SH3 leads to the underestimation of neutral density average and maximum for all models except the JB2008. For JB2008, after the SH3, the ratios approach closer to 1 with respect to other shifts for most of the events. However, there is still overestimation in two of the events. Additionally, in TIEGCM, the 2005-135 event shows a distinct behavior and captures the maximum in M2017 better after SH3. CTIPe overestimates for events with Kp<7 and underestimates in with Kp≥7 before and after the shifts.

Qualitatively, the same conclusions mostly hold true for the ratio of neutral density averages and maxima from the models; only the amount of underestimation or overestimation changes. However, a model overestimating the neutral density maximum may underestimate the average density as in JB2008 case for the 2006-348 event. Moreover, a model underestimating the neutral density maximum may overestimate the average density as in CTIPe for the event 2005-243 and GITM as in events with Kp≥7.

Timing differences between the maximum in M2017 and the models are shown on the bottom row in Figure 3. SH1, SH2 and SH3 do not change the lags between the model maximum and M2017. This is natural as only a constant value is used for the baseline shifts.

Figure 4 depicts the changes of the neutral density maximum (left panel) and average (middle panel) from the quiet time values in percentage. Right panel shows the time and orbit-integrated density change (IDC). The percentage change from the background variations and the IDC are calculated around the zero-baseline level when all climatology is removed. Accordingly, SH2 is used in the calculations of percentage change and the IDC. The percentages are calculated as $\%Change = 100 \times (storm - quiet)/quiet$..

In M2017, the change in neutral density maximum due to the geomagnetic storm is found to be nearly as twice as the change in neutral density average for the observations and models for all events. The change in neutral density maximum ranges from 200% to 90% and the change in neutral density average ranges from 100% to 45%. Both the change in maximum and average of the observations (M2017 and SV2.3) show a decreasing trend with lower geomagnetic storm intensity in terms of Kp. TIEGCM and CTIPe estimate the closest percentages to M2017 for events with Kp≤7. CTIPe also performs reasonably well for events with Kp≥7.

In the right panel, geomagnetic storms with less Kp, which are due to HSSs (2007-142, 2005-190) display IDCs as large as the events due to CMEs (2005-135, 2006-348). There is not any model, which is consistently closer to the IDC from M2017. However, MSIS is closer to M2017 more times than the other models (4 of the 6 selected cases: 2006-348, 2005-243, 2007-142, 2007-91). TIEGCM overestimates in all events. Similar to TIEGCM, JB2008 and CTIPe are higher than the M2017, except the 2005-135 and 2005-243 events, respectively. GITM shows a distinction between Kp≥6+ and Kp<6+ events: it overpredicts the IDC in events with Kp≥6+ and under predicts for events with Kp<6+ for the selected events.

Figure 5 presents mean absolute error (MAE), normalized root mean square error (NRMSE) and prediction efficiency (PE) of the models for the selected events. MAE and NRMSE are negatively-oriented skill scores, meanwhile PE is positively-oriented. This means that, lower values of MAE and NRMSE are more desirable whereas PE closer to one shows the perfect agreement between the models and M2017, in our case.

From Figure 5, the effect of baseline shifts on the storm-time performance of the models can be distinguished. It is found that generally, the calculated errors after the baseline shifts are on the same order for all models and range between 1% and 20%. In the figure, baseline shifts are found to reduce the errors (MAE, NRMSE and PE) for the TIEGCM and SV2.3 for all cases. Additionally, as in the case of the ratios, SV2.3 errors are more efficiently reduced using the SH3 compared to the other shifts.

The MAE provides information on the amount of mean error in dimensioned units (kg/m^3, in the case of thermospheric neutral density). For the selected event set, MAE is found to be high for strong events and low for weak events after the baseline shifts (except for GITM in 2005-243 and CTIPe in 2006-348), which is consistent with the findings of (Shim et al., 2012). Moreover, Figure S8 shows that the behavior of RMSE is the same with MAE in all cases and models and the amount of error grows with respect to event intensity. The amount of error increases with stronger events because the temporal variability of the thermospheric neutral density is higher in stronger geomagnetic storms. Normalization shows the errors are actually around the same order of magnitude in terms of percentage for the events. A high MAE may account for a low NRMSE based on the variation of the thermospheric neutral density during the event. On the other hand, an increase in MAE after the shift, with respect to the original time series without shift, mirrors itself as an increase in NRMSE with respect to the original time series, as well. Thus, basically, the MAE and NRMSE provide the same information on the change in errors. However, NRMSE gives the additional information that how much this error accounts for from the perspective of the variability of the thermospheric neutral density based on the event.

The NRMSE from the models are confined between 60% and 10% after the shifts. Before the shift, TIEGCM has the maximum NRMSE with ~125% for the event 2007-91. The shifts revealed that its actual storm-time performance to be on the order of ~12% (SH1, SH2) to ~33% (SH3) for the same event. In contrast, MSIS has a minimum error around ~25%, which increases to ~37% (SH1, SH2) to ~41% (SH3) for the event 2005-135.

The 2005-135 is an exceptional case as can be seen from MAE, RMSE, NRMSE and PE of the TIEGCM. Interestingly, only for TIEGCM among the other empirical and physics-based models of the IT and only in this event, baseline removal via ratios (SH3) reduces the error more than the shifts based on subtraction. In this case, the storm-time variation is so high and strong that it is compensated by taking ratios. However, we argue that this is not the actual performance of the model. Since storm effects are generally additions to the background neutral density (Lühr et al., 2011), in the case of this event, the model, in fact, overestimates the storm-time variations so much that the error is reduced via the SH3, which uses quiet-time ratios. On the other hand, for other events, SH3 gives rise to the underestimation of the average and maximum values of the neutral density from TIEGCM (Figures 2 and 3, Figures S2 to S6).

The PE on the right column shows the same variations with the NRMSE according to the event. The PE increases when the NRMSE increases and vice versa. The PEs of TIEGCM for the original, unshifted model neutral density are so low that the scales are compressed in the figure. However, after the shifts, there is a clear improvement in model performances, which can be seen from the frame interior to the figure.

The errors in TIEGCM seem to increase with the intensity of the geomagnetic storm. After removing the climatology via the baseline shifts, the errors in CTIPe and JB2008 are also found to decrease except the 2006-348 and 2005-135 events, respectively. The events with the most errors in CTIPe model are found to be the problem cases, which Knipp et al., 2013 listed (2005-190, 2005-243 and 2005-135). In the problem events, the damping of the thermospheric density by NO cooling is more than expected, so that the density may not enhance as high as, that estimated by the IT models. However, we should note that the version of CTIPe that is used in this work does not include the correction to NO cooling at high Kp levels. From the selected events, GITM appears to show a reduction in error for events with Kp≥6+ and growth in errors for the events with Kp<6 after the baseline shifts. On the contrary, after the removal of the climatology, for MSIS, the errors in the selected cases give the impression that they increase for events with Kp≥6+, except the 2005-135 event, and decrease for events with Kp<6.

In our selected cases, after SH1 or SH2, TIEGCM performed the best for events with Kp<7 according to all metrics. Moreover, TIEGCM demonstrated the highest PE for most of the cases.

Lastly, it is found that the SH3 reduces the errors more than the other shifts for the SV2.3, since neutral density is derived from the accelerometer on spacecraft and the error can be multiplied during this process. All shifts and all events in terms of MAE, NRMSE, and PE show that the SH3 works perfectly for the SV2.3 and the errors are on the order of ~1%, with a maximum of ~2.5%.

In addition to the errors from the models using the shifts SH1, SH2, and SH3 provided above, the errors for the shifts with point-to-point subtraction and multiplication (SH4, SH5, SH6) are given in the supplement (Figure S10, Figure S11). It can be seen from Figure S10 and Figure S11 that the choice of the baseline shifting method does not affect the performance outcome of the models. The errors obtained by using SH1 and SH4, SH2 and SH5 and SH3 and SH6 are very close to each other.

Furthermore, additional metrics may be utilized serving to the special purposes of the studies. For example, since their technique for data assimilation aims to reduce the errors in logarithmic densities, Sutton (2018) used mean, normalized standard deviation and root mean square (rms) errors of the log density ratio ($ln(\rho_{model}/\rho_{observation})$) in their work. For the sake of comparison, we also tested these metrics for our events. Figure S9 presents the results. The logarithmic mean gives similar results to the Ratio$_{\text{avg}}$ for all shifts, whereas for SH2, the errors from the models are amplified in standard deviation and rms relying on the *ln*(model/observation) ratio. The rms of log density ratio in SH1 and SH3 are found to be very close to NRMSE.

4 Summary and Conclusion

In this study, we had two aims: 1) to find methods to facilitate the evaluation of the storm-time performance of models and 2) to suggest a standard set of metrics to determine the model performances.

For the first part, we presented methods to remove the quiet-time bias/climatology from the models and referred to these methods as “baseline shifts”. Shifts are based on subtraction of bias from the models (SH1), subtraction of climatology from model and observation (SH2) and multiplication of the quiet-time ratio between the model and observation with the model to match the quiet-time neutral density level of observation (SH3). It was shown that defining the quiet-time reference level is very critical in determining the actual storm-time performances. In some events and models, the shifts were found to reduce the errors due to climatology in evaluating the storm-time performances up to 113% (TIEGCM-2007-91: 125% to 12%) whereas in some events, they increased the errors by 13% (MSIS-2005-135: 12% to 25%).

For the storm-time performance assessment of the models, SH1 and SH2 are found to work equally well. The choice of different baseline levels (shifting the models to the level of CHAMP observations or shifting observations and models to zero level by removing all the climatology) does not change the amount of error associated with a model. Besides, SH3 increases the variability of the errors from the models when compared to the other shifts. This is due to the fact that the storm-time effects are generally superimposed upon the background (climatological) variations and their nature is not multiplicative. Hence, modifying the original time series using ratios does not work as efficiently as the subtraction process for the empirical and physics-based models.

On the other hand, SH3 is efficient when comparing M2017 and SV2.3 as it depends on the quiet time ratios. The difference between these two data sets is only a constant number, which depends on the modeling of the C_d and the geometry of the spacecraft. Therefore, the SH3 works the best for SV2.3 when compared to other shifts. Hence, when neutral density is derived from accelerometer data, systematic error and bias can be multiplied, so it is reasonable to divide to remove them. It follows that the findings of the past model validation studies which used SV2.3 can be re-evaluated and calibrated using the SH3.

From the selected cases, it appears that, TIEGCM is more successful in low Kp events, and its success rate decreases with the intensity of the storm. GITM shows a reduction in error for events with Kp≥6+ and increase in errors for the events with Kp<6. On the contrary, the model errors increase for MSIS for events with Kp≥6+, except the 2005-135 event, and decrease for events with Kp<6 in this event set. JB2008 does not show any systematic errors for the selected events. After the removal of the quiet time bias/climatology between M2017 and the models, TIEGCM seems to perform the best in terms of all metrics for most of the selected events, followed by CTIPe and GITM. For the selected cases, JB2008 was closer to M2017 than MSIS for more of the events.

Three of the six events selected in this study were listed as problem storms by (Knipp et al., 2013). They reported that the modeling of these storms is more difficult with respect to several other events with less NO production. The NO cooling during these events restrict the neutral density enhancement and neutral density does not increase as high as expected from the models. In our event set, for these storms, the range of errors from the models are between 13% and 40%

and do not greatly differ from the other cases. Thus, we do not see any distinction among model performances with respect to the solar wind drivers of the events. On the other hand, performances of the MSIS, TIEGCM and GITM suggest differences based on Kp.

Furthermore, it is possible to estimate integrated neutral density change (IDC) during the storm via SH2, which shifts the baseline to zero level. IDC is important as drag has a cumulative effect on orbit determination and prediction (Emmert et al., 2017). For the evaluations in drag calculations, we suggest using the upper limits for IDC that are calculated after SH2 to stay on the safe side. In terms of the IDC metric, MSIS was found to be the closest to the M2017 in more events than the other models studied here. This may be due to the fact that MSIS is trained with the integrated neutral density (Picone et al., 2002).

The second part of this study involves selecting a standard set of metrics to quantify the errors in neutral density. Seven metrics were investigated for this purpose: the ratio between the model maximum and CHAMP maximum (Ratio$_{max}$), ratio between the model mean and CHAMP mean (Ratio$_{avg}$), time delay between the peak of the model and peak of the CHAMP observation (TD), mean absolute error (MAE), normalized root mean square error (NRMSE), prediction efficiency (PE) and integrated density change (IDC). In this study, we show that Ratio$_{max}$ and Ratio$_{avg}$ may not be consistent with each other even after the baseline shifting procedure. A model overestimating the ratio of maximum may predict the Ratio$_{avg}$ well. This is due to the shape of the response curve and is controlled by how fast the growth and decay rates of the neutral density are within the model. Thus, neither the neutral density maximum nor the neutral density average is definitive in model performance assessment when used alone. In this study, consistency is achieved between the skill scores MAE, RMSE, NRMSE and PE after the baseline shifts. Consequently, we suggest using MAE, NRMSE and PE together for the neutral density evaluations. MAE will provide the mean amount of error, NRMSE, the error percentage with respect to the event and PE will provide how efficient the model is in capturing the variability and mean of the neutral density observations.

To conclude, we have shown that baseline shifting is useful in assessing the storm-time model performance when models have bias against the data during the quiet-time. Removal of the baseline allows for the detection of actual storm-time response and performances from the models. We emphasize that quiet-time climatology and storm-time performances of the models should be evaluated separately, and after baseline shifts, especially for the models with quiet-time bias. Even though, we focused on the storm-time performances of the models in this work, we emphasize that for the long-term estimations of satellite drag, it is important to provide the background neutral density precisely.

For satellite drag calculations, the accuracy of neutral density estimations is important. This study shows, the IT models present variable errors depending on the event. None of the models perform perfectly for all cases. In such cases, the uncertainty in thermospheric neutral density in an event can be represented well by using an ensemble of models and iterating the results (Elvidge et al., 2016). In an operational scenario, the ensemble method and baseline shifts using the previous, quiet-day estimations can be used together to tune the models and their output, so that the storm-time variations can be better estimated. Murray (2018) demonstrated the usefulness of ensembles in space weather forecasting to determine the uncertainty and (Knipp, 2016) reported the studies, which use the ensemble method for space weather forecasting. We also point out that multi-model ensemble forecasts can be of great use and are candidates for future work, especially in respect of the integrated density change, maximum and average neutral density which are found to be highly variable among the models and are important in satellite drag calculations and for real-time operations.

Acknowledgments

This work has been supported by the Turkish Scientific and Technological Council (TÜBİTAK), Project No: 113Y213 and 2214/A-International Doctoral Research Fellowship Programme. We thank the producers of the Dst index at WDC, Kyoto (http://wdc.kugi.kyoto-u.ac.jp/) and GFZ, Potsdam for the Kp index values (ftp://ftp.ngdc.noaa.gov/STP/GEOMAGNETIC_DATA/INDICES/Kp_AP/). HP index values were obtained from cedarweb (https://cedarweb.vsp.ucar.edu/wiki/index.php/Tools_and_Models:Emery_HP_plus_indices_to_11107). F10.7 values were accessed from (ftp://ftp.ngdc.noaa.gov/STP/space-weather/solar-data/solar-features/solar-radio/noontime-flux/penticton/penticton_observed/listings/listing_drao_noontime-flux-observed_daily.txt). We thank all of the abovementioned for giving open access to the data. Neutral density data were obtained from (https://drive.google.com/drive/folders/0BwtX8XEH-aEueHJiU1htLVo0cms?usp=drive_open) provided by (Mehta et al., 2017). Model runs and results are made available through the NASA Community Coordinated Modeling Center (CCMC) through their public Runs on Request system (http://ccmc.gsfc.nasa.gov). Model results for the individual simulations can be searched from the CCMC-View Results with the Simulation IDs listed in Table S1. The CCMC is a multi-agency partnership between NASA, AFMC, AFOSR, AFRL, AFWA, NOAA, NSF and ONR. Lastly, we thank the reviewers and especially editor Michael A. Hapgood for their valuable comments, which improved our paper.

References

Anderson, R. L., Born, G. H., & Forbes, J. M. (2009). Sensitivity of Orbit Predictions to Density Variability. Journal of Spacecraft and Rockets, 46(6), 1214–1230. https://doi.org/10.2514/1.42138

Bowman, B. R., Tobiska, W. K., Marcos, F. A., Huang, C. Y., Lin, C. S., & Burke, W. J. (2008). 37 TH COSPAR Scientific Assembly 2008 – Proposal for CIRA 2008, A New Empirical Thermospheric Density Model JB2008 Using New Solar and Geomagnetic Indices, (August).

Bruinsma, S. L., & Forbes, J. M. (2010). Anomalous behavior of the thermosphere during solar minimum observed by CHAMP and GRACE. Journal of Geophysical Research: Space Physics, 115(11), 1–8. https://doi.org/10.1029/2010JA015605.

Bruinsma, S., Sutton, E., Solomon, S. C., Fuller -Rowell, T., & Fedr weather modeling capabilities assessment: Neutral density for orbit determination at low Earth orbit. Space Weather, 16. https://doi.org/10.1029/2018SW002027.

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

Bussy-Virat, C. D., Ridley, A. J., & Getchius, J. W. (2018). Effects of Uncertainties in the Atmospheric Density on the Probability of Collision Between Space Objects. Space Weather, 519–537. https://doi.org/10.1029/2017SW001705

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, G., Xu, J., Wang, W., Lei, J., & Burns, A. G. (2012). A comparison of the effects of CIR- and CME-induced geomagnetic activity on thermospheric densities and spacecraft orbits: Case studies. Journal of Geophysical Research: Space Physics, 117(A8). https://doi.org/10.1029/2012JA017782

Chen, G., Xu, J., Wang, W., & Burns, A. G. (2014). A comparison of the effects of CIR- and CME-induced geomagnetic activity on thermospheric densities and spacecraft orbits: Statistical studies. Journal of Geophysical Research: Space Physics, 119(9), 7928–7939. https://doi.org/10.1002/2014JA019831

Codrescu, M. V., Fuller-Rowell, T. J., Munteanu, V., Minter, C. F., & Millward, G. H. (2008). Validation of the coupled thermosphere ionosphere plasmasphere electrodynamics model: CTIPE-mass spectrometer incoherent scatter temperature comparison. Space Weather, 6(9), 1–10. https://doi.org/10.1029/2007SW000364

Codrescu, M. V., Negrea, C., Fedrizzi, M., Fuller-Rowell, T. J., Dobin, A., Jakowsky, N., … Maruyama, N. (2012). A real-time run of the Coupled Thermosphere Ionosphere Plasmasphere Electrodynamics (CTIPe) model. Space Weather, 10(1), 1–10. https://doi.org/10.1029/2011SW000736

Connor, H. K., Zesta, E., Fedrizzi, M., Shi, Y., Raeder, J., Codrescu, M. V., & Fuller-Rowell, T. J. (2016). Modeling the ionosphere-thermosphere response to a geomagnetic storm using physics-based magnetospheric energy input: OpenGGCM-CTIM results. Journal of Space Weather and Space Climate, 6, A25. https://doi.org/10.1051/swsc/2016019

Deng, Y., Fuller-Rowell, T. J., Ridley, A. J., Knipp, D., & Lopez, R. E. (2013). Theoretical study: Influence of different energy sources on the cusp neutral density enhancement. Journal of Geophysical Research: Space Physics, 118(5), 2340–2349. https://doi.org/10.1002/jgra.50197

Elvidge, S., Angling, M. J., & Nava, B. (2014). On the use of modified Taylor diagrams to compare ionospheric assimilation models. Radio Science, 49(9), 737–745. https://doi.org/10.1002/2014RS005435

Elvidge, S., Godinez, H. C., & Angling, M. J. (2016). Improved forecasting of thermospheric densities using multi-model ensembles. Geoscientific Model Development, 9(6), 2279–2292. https://doi.org/10.5194/gmd-9-2279-2016

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

Fedrizzi, M., Fuller-Rowell, T. J., & Codrescu, M. V. (2012). Global Joule heating index derived from thermospheric density physics-based modeling and observations. Space Weather, 10(3). http://doi.org/10.1029/2011SW000724

Hejduk, M. D., & Snow, D. E. (2018). The Effect of Neutral Density Estimation Errors on Satellite Conjunction Serious Event Rates. Space Weather. https://doi.org/10.1029/2017SW001720

Huang, C. Y., Su, Y.-J., Sutton, E. K., Weimer, D. R., & Davidson, R. L. (2014). Energy coupling during the August 2011 magnetic storm. Journal of Geophysical Research: Space Physics, 119(2), 1219–1232. https://doi.org/10.1002/2013JA019297

Kim, K. H., Moon, Y. J., Cho, K. S., Kim, H. D., & Park, J. Y. (2006). Atmospheric drag effects on the KOMPSAT-1 satellite during geomagnetic superstorms. Earth, Planets and Space, 58(12), 25–28. https://doi.org/10.1186/BF03351968

Knipp, D., Kilcommons, L., Hunt, L., Mlynczak, M., Pilipenko, V., Bowman, B., & Deng, Y. (2013). Thermospheric damping response to sheath-enhanced geospace storms. Geophysical Research Letters, 40(7), 1263–1267. https://doi.org/10.1002/grl.50197

Knipp, D. J. (2016), Advances in Space Weather Ensemble Forecasting, Space Weather, 14, 52–53, doi:10.1002/2016SW001366.

Kwak, Y. S., Richmond, A. D., Deng, Y., Forbes, J. M., & Kim, K. H. (2009). Dependence of the high-latitude thermospheric densities on the interplanetary magnetic field. Journal of Geophysical Research: Space Physics, 114(5), 1–7. https://doi.org/10.1029/2008JA013882

Lathuillère, C., Menvielle, M., Marchaudon, A., & Bruinsma, S. (2008). A statistical study of the observed and modeled global thermosphere response to magnetic activity at middle and low latitudes. Journal of Geophysical Research: Space Physics, 113(7), 1–9. https://doi.org/10.1029/2007JA012991

Lei, J., Thayer, J. P., Lu, G., Burns, A. G., Wang, W., Sutton, E. K., & Emery, B. A. (2011). Rapid recovery of thermosphere density during the October 2003 geomagnetic storms. Journal of Geophysical Research: Space Physics, 116(3), 1–10. https://doi.org/10.1029/2010JA016164

Liu, H., H. Lühr, V. Henize, and W. Köhler (2005), Global distribution of the thermospheric total mass density derived from CHAMP, J. Geophys. Res.,110, A04301, doi:10.1029/2004JA010741.

Lühr, H., Rother, M., Köhler, W., Ritter, P., & Grunwaldt, L. (2004). Thermospheric up-welling in the cusp region: Evidence from CHAMP observations. Geophysical Research Letters, 31(6). https://doi.org/10.1029/2003GL019314

Lühr, H., Liu, H., Park, J., & Müller, S. (2011). New Aspects of the Coupling Between Thermosphere and Ionosphere, with Special regards to CHAMP Mission Results. In M. A.

Abdu & D. Pancheva (Eds.), Aeronomy of the Earth’s Atmosphere and Ionosphere (pp. 303–316). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0326-1

McGranaghan, R., Knipp, D. J., McPherron, R. L., & Hunt, L. A. (2014). Impact of equinoctial high-speed stream structures on thermospheric responses. Space Weather, 12(4), 277–297. https://doi.org/10.1002/2014SW001045

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

Millward, G. H., Müller-Wodarg, I. C. F., Aylward, A. D., Fuller-Rowell, T. J., Richmond, A. D., & Moffett, R. J. (2001). An investigation into the influence of tidal forcing on F region equatorial vertical ion drift using a global ionosphere-thermosphere model with coupled electrodynamics. Journal of Geophysical Research, 106(A11), 24733. https://doi.org/10.1029/2000JA000342

Murray, S. A. (2018). The importance of ensemble techniques for operational space weather forecasting. Space Weather, 1–10. https://doi.org/10.1029/2018SW001861

Pardini, C., Moe, K., & Anselmo, L. (2012). Thermospheric density model biases at the 23rd sunspot maximum. Planetary and Space Science, 67(1), 130–146. https://doi.org/10.1016/j.pss.2012.03.004

Picone, J. M., Hedin, A. E., Drob, D. P., & Aikin, A. C. (2002). NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. Journal of Geophysical Research: Space Physics, 107(A12), SIA 15-1-SIA 15-16. https://doi.org/10.1029/2002JA009430

Prölss, G. W. (2011). Density Perturbations in the Upper Atmosphere Caused by the Dissipation of Solar Wind Energy. Surveys in Geophysics, 32(2), 101–195. https://doi.org/10.1007/s10712-010-9104-0

Qian, L., Solomon, S. C., Roble, R. G., Bowman, B. R., & Marcos, F. A. (2008). Thermospheric neutral density response to solar forcing. Advances in Space Research, 42(5), 926–932. https://doi.org/10.1016/j.asr.2007.10.019

Qian, L., & Solomon, S. C. (2012). Thermospheric Density: An Overview of Temporal and Spatial Variations. Space Science Reviews, 168(1–4), 147–173. https://doi.org/10.1007/s11214-011-9810-z.

Rhoden, E. A., Forbes, J. M., & Marcos, F. A. (2000). The influence of geomagnetic and solar variabilities on lower thermosphere density. Journal of Atmospheric and Solar-Terrestrial Physics, 62(11), 999–1013. https://doi.org/10.1016/S1364-6826(00)00066-3.

Richmond, A. D., Ridley, E. C., & Roble, R. G. (1992). A thermosphere/ionosphere general circulation model with coupled electrodynamics. Geophysical Research Letters, 19(6), 601–604. https://doi.org/10.1029/92GL00401

Ridley, A. J., Deng, Y., & Tóth, G. (2006). The global ionosphere–thermosphere model. Journal of Atmospheric and Solar-Terrestrial Physics, 68(8), 839–864. https://doi.org/10.1016/j.jastp.2006.01.008

Shim, J. S., et al. (2011), CEDAR Electrodynamics Thermosphere Ionosphere (ETI) Challenge for systematic assessment of ionosphere/thermosphere models: NmF2, hmF2, and vertical drift using ground-based observations, Space Weather, 9, S12003, doi:10.1029/2011SW000727.

Shim, J. S., Kuznetsova, M., Rastätter, L., Bilitza, D., Butala, M., Codrescu, M., … Sutton, E. (2012). CEDAR Electrodynamics Thermosphere Ionosphere (ETI) Challenge for systematic assessment of ionosphere/thermosphere models: Electron density, neutral density, NmF2, and hmF2 using space-based observations. Space Weather, 10(10). https://doi.org/10.1029/2012SW000851

Shim, J. S., Kuznetsova, M., Rastätter, L., Bilitza, D., Butala, M., Codrescu, M., … Sutton, E. (2014). Systematic Evaluation of Ionosphere/Thermosphere (IT) Models (pp. 145–160). American Geophysical Union (AGU). https://doi.org/10.1002/9781118704417.ch13

Shim, J. S., L. Rastaetter, K. M. Kuznetsova, E. C. Kalafatoglu, and Y. Zheng (2015). Assessment of the predictive capability of IT models at the Community Coordinated Modeling Center. Presented at Ionospheric Effect Symposium, Alexandria VA.

Solomon, S. C., Qian, L., Didkovsky, L. V., Viereck, R. A., & Woods, T. N. (2011). Causes of low thermospheric density during the 2007-2009 solar minimum. Journal of Geophysical Research: Space Physics, 116(7), 1–14. https://doi.org/10.1029/2011JA016508

Storz, M. F., Bowman, B. R., Branson,M. J. I., Casali, S. J., & Tobiska,W. K. (2005). High accuracy satellite drag model (HASDSM). Advances in Space Research, 36(12), 2497–2505.

Sutton, E. K. (2008), Effects of solar disturbances on the thermosphere densities and winds from CHAMP and GRACE satellite accelerometer data, Doctoral Dissertation, Dept. of Aerosp. Eng. Sci., Univ. of Colorado, Boulder, Colo.

Sutton, E. K. (2009), Normalized force coefficients for satellites with elongated shapes, J. Spacecraft and Rockets, 46(1), doi:10.2514/1.40940.

Sutton, E. K. (2011). Accelerometer-Derived Atmospheric Density from the CHAMP and GRACE Satellites. Version 2.3. AIR FORCE RESEARCH LAB KIRTLAND AFB NM.

Sutton, E. K. (2018). A New Method of Physics-Based Data Assimilation for the Quiet and Disturbed Thermosphere. Space Weather, 16(6), 736–753. https://doi.org/10.1002/2017SW001785

Sutton, E. K., Forbes, J. M., & Nerem, R. S. (2005). Global thermospheric neutral density and wind response to the severe 2003 geomagnetic storms from CHAMP accelerometer data. Journal of Geophysical Research: Space Physics, 110(A9), 1–10. https://doi.org/10.1029/2004JA010985

Sutton, E. K., Forbes, J. M., Nerem, R. S., & Woods, T. N. (2006). Neutral density response to the solar flares of October and November, 2003. Geophysical Research Letters, 33(22), 1–5. https://doi.org/10.1029/2006GL027737

Thayer, J. P., Lei, J., Forbes, J. M., Sutton, E. K., & Nerem, R. S. (2008). Thermospheric density oscillations due to periodic solar wind high speed streams. Journal of Geophysical Research: Space Physics, 113(6), A06307. https://doi.org/10.1029/2008JA013190

Xu, J., Wang, W., Lei, J., Sutton, E. K., & Chen, G. (2011). The effect of periodic variations of thermospheric density on CHAMP and GRACE orbits. Journal of Geophysical Research: Space Physics, 116(2), A02315. https://doi.org/10.1029/2010JA015995

Weimer, D. R. (2005). Improved ionospheric electrodynamic models and application to calculating Joule heating rates. Journal of Geophysical Research, 110(A5), A05306. https://doi.org/10.1029/2004JA010884

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, 79–82.

Zesta, E., & Huang, C. Y. (2016). Satellite orbital drag. In G. V. Khazanov (Ed.), Space weather fundamentals (pp. 329–351). Boca Raton, FL: CRC Press.

Table 1. GEM-CEDAR Challenge events. Table shows the maximum values of geomagnetic and solar indices (Kp_{max}, F10.7, Dst_{max}, HP_{max}) and solar wind drivers of the events.

Event	**Kp_{max}**	**F10.7**	**Dst_{min} (nT)**	**HP_{max} (GW)**	**Driver**
2005-135	8+	103	-247	1225	CME
2006-348	8+	93.6	-162	504	CME
2005-243	7	84	-122	260	HSS
2005-190	6+	106.6	-92	238	HSS
2007-142	5+	72	-58	197	HSS
2007-091	5	71.7	-63	286	HSS

Table 2. Baseline shifts. ρ_{old} is the original orbit-averaged time series whereas ρ_{new} is the baseline shifted time series. Subscript index "n" represents the orbit numbers for the entire event (quiet+storm) interval, "i" stands for the orbit number during the selected quiet time interval of the event. Overbars denote the mean.

Shifts	Shifting Parameter	Shifted Series	Reference Level
Shift1 (SH1)	$S_1=\overline{\rho_{champ,\imath} - \rho_{model,\imath}}$	$\rho_{new,n=}\ \rho_{old,n}-S_1$	CHAMP
Shift2 (SH2)	$S_2=\overline{\rho_{champ,\imath}}$ for CHAMP $S_2=\overline{\rho_{models,\imath}}$ for models	$\rho_{new,n=}\ \rho_{old,n}-S_2$	Zero
Shift3 (SH3)	$S_3=\overline{\rho_{champ,\imath}/\rho_{model,\imath}};$	$\rho_{new,n=}\ \rho_{old,n}\times S_3$	CHAMP

Figures:

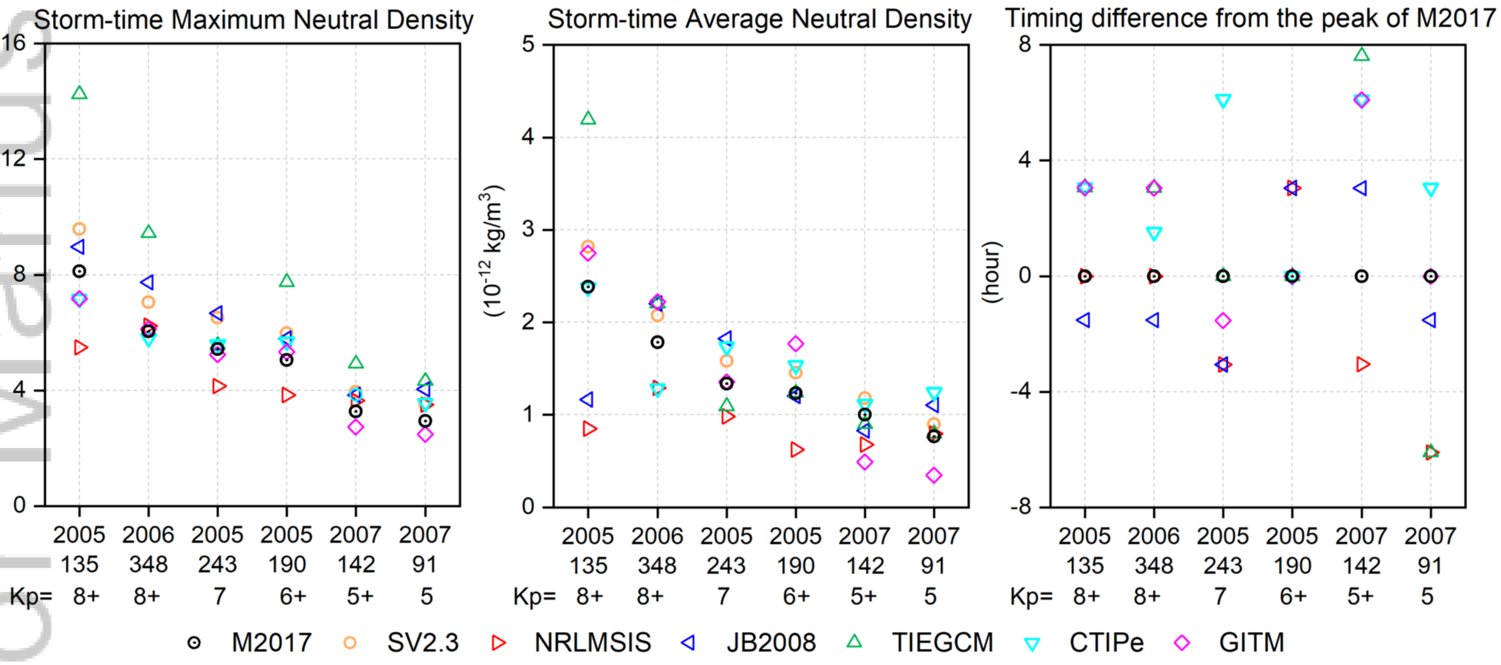
Figure 1. From left to right: storm-time maximum in neutral density, storm-time average neutral density, timing difference between the peak of models and M2017. The circles denote neutral density estimations based on accelerometers on CHAMP: orange, SV2.3 and dot-centered black, M2017. The triangles and the diamond show the model estimations: red, right-triangle: MSIS; blue, left-triangle: JB2008; green, up-triangle: TIEGCM; cyan, down-triangle: CTIPe; pink, diamond: GITM. X-label is the events listed from severe (Kp>8) to weak (Kp=5) starting from left to right, according to the NOAA classification based on Kp values.

Figure 2: An example event: 2006-348. First row, from left to right: a) top: Neutral density from the model and observations without shift; below: Kp and Dst indices, neutral density estimations from the models and M2017 after b) SH1, c) SH2, d) SH3.

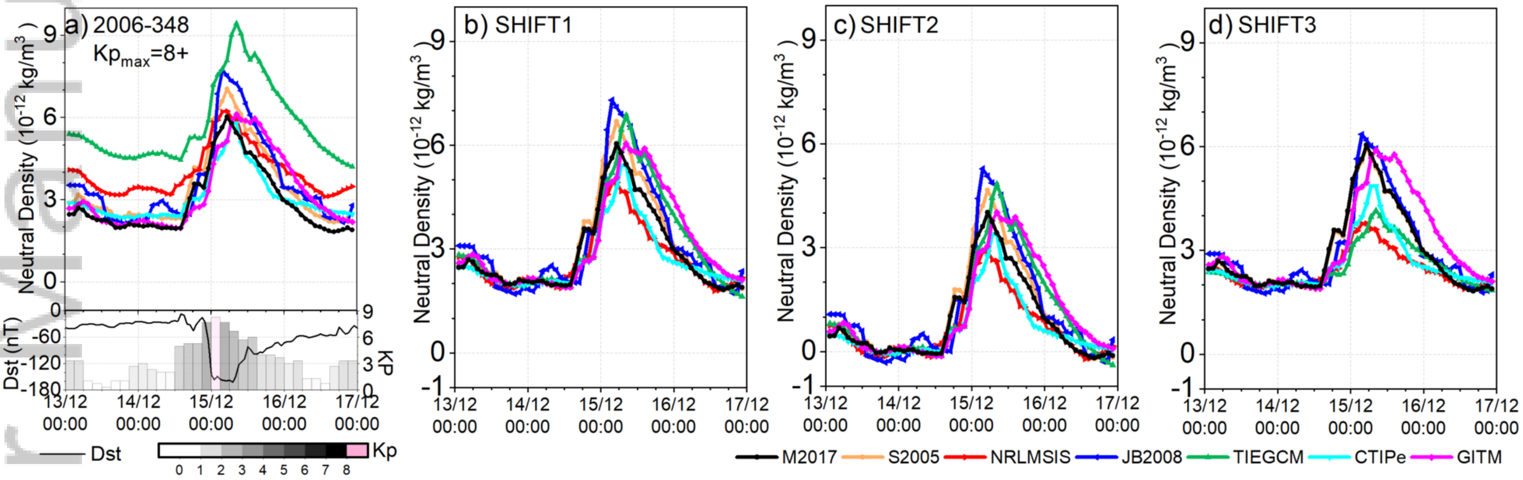
Figure 3: From top to bottom: storm-time ratio of maximum neutral density of the models to M2017, storm-time ratio of average neutral density from the models to M2017, timing difference between the peak of models and M2017. From left to right: SV2.3, MSIS, JB2008, CTIPe, GITM and TIEGCM. O denotes the results for the original, unshifted time series whereas SH1 to SH3 represents the shifts from Shift1 to Shift3. Red symbols represent the severe events with high Kp; cyan denotes strong event with Kp=7; black is for 7>Kp>6; and green color is for weak events with Kp around 5. Circle represents the event 2005-135; square, 2006-348; up-triangle, 2005-243; down-triangle, 2005-190; cross, 2007-142; plus, 2007-91.

Figure 4: From left to right: storm-time orbit and time integrated neutral density, storm-time change in maximum neutral density, storm-time change in mean neutral density. The symbol and colors are the same as Figure 1.

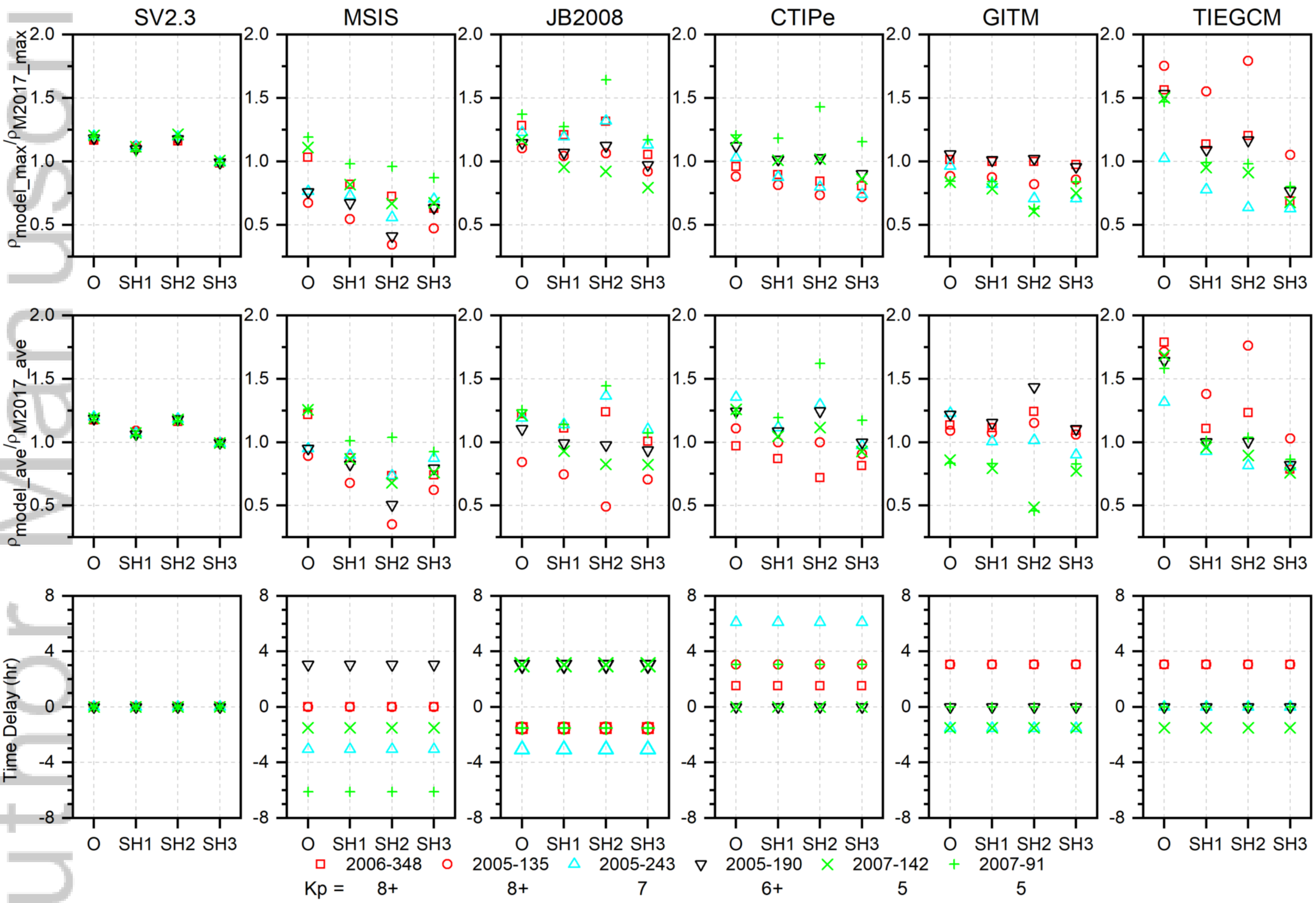
Figure 5: From top to bottom: MAE, NRMSE, PE. From left to right: SV2.3, MSIS, JB2008, CTIPe, GITM and TIEGCM. KP scales, axis labels, colors and symbols are the same as Figure 3. Please note that the y-axis scales for TIEGCM is different than the other panels for the three parameters. Additionally, for TIEGCM, PE results after the shifts SH1 to SH3 are shown in another frame inside the PE panel with scaled y-axis. The inside frame has the same y-axis scale as the other panel for PEs.



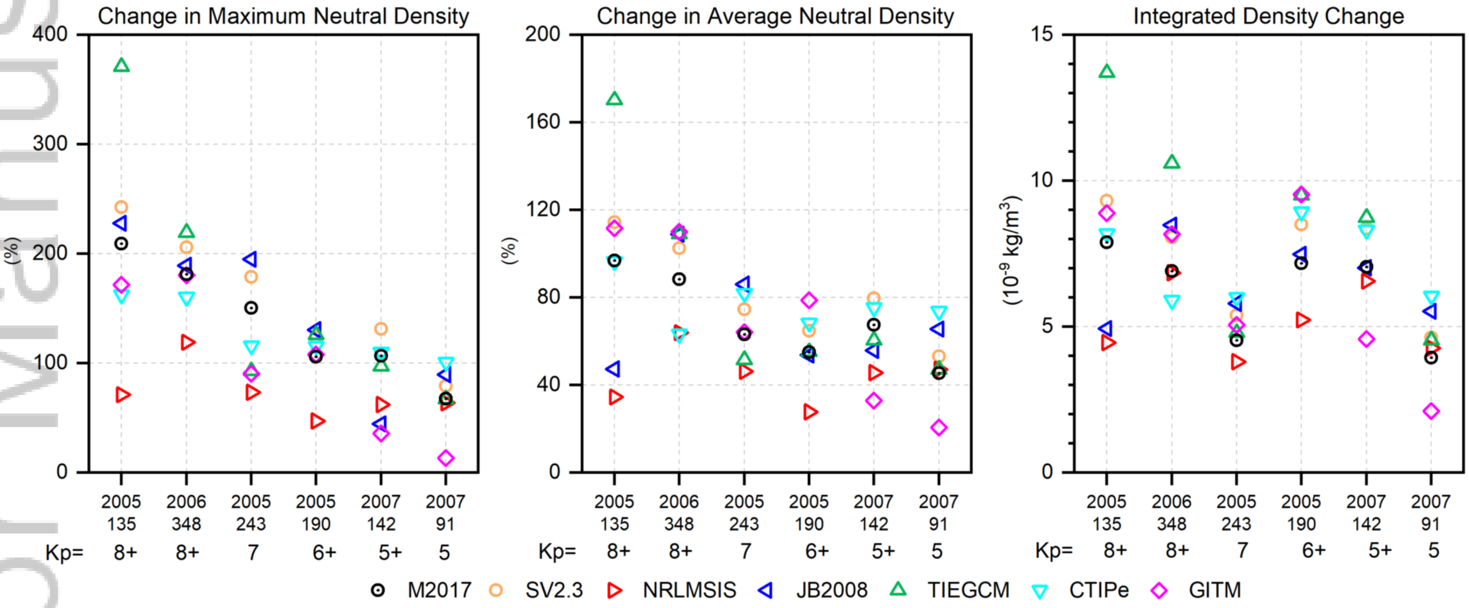
2018SW002033-f01-z-.png



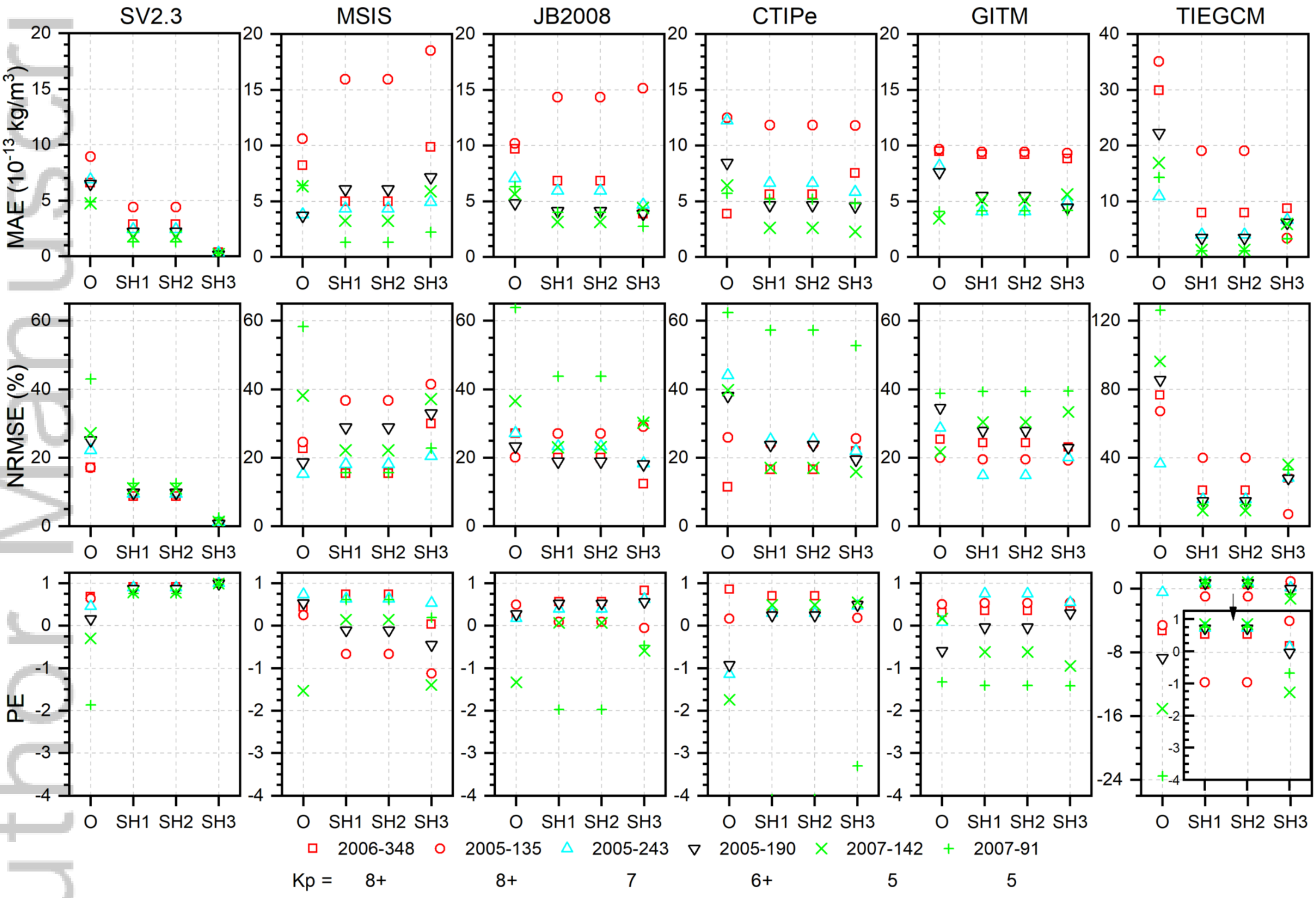
2018SW002033-f02-z-.png



2018SW002033-f03-z-.png



2018SW002033-f04-z-.png



2018SW002033-f05-z-.png