J. Space Weather Space Clim. 2021, **11**, 12

https://doi.org/10.1051/swsc/2021002

JSWSC

Available online at:
www.swsc-journal.org

RESEARCH ARTICLE OPEN ACCESS

Thermosphere modeling capabilities assessment: geomagnetic storms

Sean Bruinsma[1,*], Claude Boniface[1], Eric K. Sutton[2], and Mariangel Fedrizzi[3]

[1] CNES, Space Geodesy Office, 31401 Toulouse, France
[2] University of Colorado, Space Weather Technology, Research & Education Center (SWx TREC), Boulder, 80309-0429 CO, USA
[3] University of Colorado/CIRES and NOAA/SWPC, Boulder, 80305 CO, USA

Received 12 June 2020 / Accepted 8 January 2021

Abstract – The specification and prediction of density fluctuations in the thermosphere, especially during geomagnetic storms, is a key challenge for space weather observations and modeling. It is of great operational importance for tracking objects orbiting in near-Earth space. For low-Earth orbit, variations in neutral density represent the most important uncertainty for propagation and prediction of satellite orbits. An international conference in 2018 conducted under the auspices of the NASA Community Coordinated Modeling Center (CCMC) included a workshop on neutral density modeling, using both empirical and numerical methods, and resulted in the organization of an initial effort of model comparison and evaluation. Here, we present an updated metric for model assessment under geomagnetic storm conditions by dividing a storm in four phases with respect to the time of minimum *Dst* and then calculating the mean density ratios and standard deviations and correlations. Comparisons between three empirical (NRLMSISE-00, JB2008 and DTM2013) and two first-principles models (TIE-GCM and CTIPe) and neutral density data sets that include measurements by the CHAMP, GRACE, and GOCE satellites for 13 storms are presented. The models all show reduced performance during storms, notably much increased standard deviations, but DTM2013, JB2008 and CTIPe did not on average reveal a significant bias in the four phases of our metric. DTM2013 and TIE-GCM driven with the Weimer model achieved the best results taking the entire storm event into account, while NRLMSISE-00 systematically and significantly underestimates the storm densities. Numerical models are still catching up to empirical methods on a statistical basis, but as their drivers become more accurate and they become available at higher resolutions, they will surpass them in the foreseeable future.

Keywords: data and metrics for standardized thermosphere model assessment / assessments for TIE-GCM, CTIPe, NRLMSISE-00, JB2008 and DTM2013 / short descriptions are given for each of the models

1 Introduction

Thermosphere models are used operationally mainly in the determination and prediction of orbits of active satellites and orbital debris, and conjunction analysis is becoming a major issue with the fast-growing number of objects in space. The accuracy of the determination and prediction of ephemerides of objects in Low Earth Orbit (LEO; altitudes lower than 1000 km) hinges on the quality of the force model for atmospheric drag (Hejduk & Snow, 2018). This force depends, besides on satellite characteristics (Doornbos, 2011; Mehta et al., 2017), heavily on the highly variable, both spatially as well as temporally, total neutral density, and to a lesser degree also on composition and temperature. Thermosphere variability is driven, on different time scales, by the changing solar extreme UV (EUV) emissions, Joule heating and particle precipitation due to interaction of the magnetosphere with the solar wind (referred to as "geomagnetic activity" as opposed to "solar activity"), and due to upward propagating perturbations that originate in Earth's lower atmosphere, which are currently not accurately quantified (Pedatella et al., 2014). As a result, the selected solar and geomagnetic activity drivers, and in case of density prediction, the accuracies of their forecasts, are crucial to thermosphere model accuracy. The impact due to errors in the driver forecasts is out of the scope of this paper but has recently been addressed by Bussy-Virat et al. (2018) and Hejduk & Snow (2018).

*Corresponding author: sean.bruinsma@cnes.fr

In order to track the progress over time of thermosphere first principle (FP) and semi-empirical (SE) models, appropriate metrics are required. Secondly, high quality neutral density data sets, preferably with high-spatial resolution covering long intervals of time (i.e. years, and ideally a complete solar cycle), are needed. The neutral density observations can then be used to verify model accuracy in space and time, i.e. with respect to latitude–longitude-local time variations, and solar and geomagnetic activity levels and seasonal variations. Data and metrics allow benchmarking of the models, quantifying errors and performance, and giving detailed descriptions of the improved (or degraded) performance.

Metrics and results for thermosphere model assessment on timescales from years to days have already been discussed and published (Bruinsma et al., 2018), but the metrics are not well-suited to describe model performance during geomagnetic storms. This is due to the relative rareness of strong storms (average frequency of 60–200 days per 11-year cycle; https://www.swpc.noaa.gov/noaa-scales-explanation), but also due to their sudden occurrences and relatively short durations (1–3 days typically). We propose specific additional metrics for storm-time, using high-quality and high-resolution density data that is required for model comparisons under storm conditions, using most of the events defined in Bruinsma et al. (2018). The assessment procedure and metrics are tailored to geomagnetic storms by unambiguously defining the time interval of evaluation, applying the same metrics as in the Bruinsma et al. (2018) study but on 4 specific phases of the storm as well as on the total interval. Secondly, additional metrics concern the maximum and timing of the storm peak density.

Presently, evaluations of both SE and FP models are available for single storms (Forbes et al., 1987, 2005; Bruinsma et al., 2006) or several storms (Liu & Luehr, 2005; Knipp et al., 2017), or data of a specific satellite mission (Kalafatoglu Eyiguler et al., 2019). Because different metrics, data, or event durations were used, results are not directly comparable even in the case that the same storm was analyzed. The ultimate goal of this exercise is to evaluate all thermosphere models available on the CCMC (Community Coordinated Modeling Center: https://ccmc.gsfc.nasa.gov) by comparing to the same data and for the same events, applying consistent and always identical metrics, in order to establish score cards that can help users select the best model for their objective.

The same three SE models, NRLMSISE-00 (Picone et al., 2002), JB2008 (Bowman et al., 2008) and DTM2013 (Bruinsma, 2015), and FP models, TIE-GCM (Roble et al., 1988; Richmond et al., 1992) and CTIPe (Fuller-Rowell et al., 1996), which were used in the Bruinsma et al. (2018) assessment, are considered in this paper. These five models are implemented at CCMC. The model resolution and drivers used in this assessment are listed in Table 1. One has to take into account that JB2008 regularly changes the files with the solar drivers, which are computed and modified by the modelers. The driver files (SOLFSMY and DTCFILE) that were online in the month March 2019 were used for the assessment given in this paper.

Section 2 provides short descriptions of the five models tested in this first part of the assessment. Section 3 presents the storm-time metrics designed to assess the models, which is applied to the five models in Section 4. After a short summary, the conclusions are given in Section 5.

2 Model descriptions

The following three sections are rather similar to the descriptions in (Bruinsma et al., 2018), but are repeated in this paper to be self-contained and for ease of reading.

2.1 Semi-empirical thermosphere models: NRLMSISE-00, JB2008 and DTM2013

SE models are mainly used in orbit computation and mission design and planning, and sometimes to provide initial conditions for FP models. They are easy to use and computationally fast, providing density and temperature estimates for a single location at a time (e.g., for each orbit position). They are climatology (or "specification") models that have a low spatial resolution of the order of thousands of kilometers, which is due to the low maximum degree (typically < 6) of the spherical harmonic expansion used in the algorithm, and low temporal resolution of hours, imposed by the cadence of the geomagnetic indices. Consequently, SE models cannot reproduce the realistic wave-like activity during geomagnetic storms such as large-scale traveling atmospheric disturbances (TAD; Bruinsma & Forbes, 2007), the complex dynamics in the polar caps, or the time-variable effects of tidal perturbations propagating from the lower atmosphere. They also do not take into account any storm pre-conditioning; the estimate is entirely based on a combination of statistical fits to the driver inputs. The minimum altitude of JB2008 and DTM2013 is 120 km, whereas for NRLMSISE-00 it is 0 km, and each can be used to approximately 1500 km.

SE models are constructed by optimally estimating the model coefficients to data in a least-squares sense. Each model is based on a different combination of density, temperature, and composition measurements. The main sources of density data are satellite-drag inferred total densities by means of orbit perturbation analysis (Jacchia & Slowey, 1963) or accelerometers (e.g. Champion & Marcos, 1973), and neutral mass spectrometers (e.g. Nier et al., 1973), with the latter providing composition measurements. Drag-inferred and spectrometer data have in common that they do not provide an absolute measurement of density. This is due to calibration issues and, e.g., unknown oxygen recombination rates in case of mass spectrometers. The most recent and precise accelerometer-inferred density datasets are also not absolute. Their magnitude depends on the satellite model, and in particular the aerodynamic coefficient, which effectively is a scaling factor, that was assumed in the computation. As a consequence, the SE thermosphere models fit to the scales of the satellite models used in the respective databases – and these are rarely consistent between modelers. Intercalibration is a necessary and complicated activity for all modelers, but it is not always entirely successful due to e.g., no overlap in time or a drifting offset between datasets. JB2008 (up to 2008 at least) and DTM2013 predict densities that are close (often within 5%) to the US Air Force operational thermosphere model HASDM (Storz et al., 2005), and therefore they are considered having the same scale.

2.2 TIE-GCM

The NCAR Thermosphere–Ionosphere–Electrodynamics General Circulation model (TIE-GCM) is a first-principles upper atmospheric general circulation model that solves the

Table 1. Thermosphere and upper atmosphere models used in the assessment.

	Type	Drivers (solar | geomagnetic)	Hor. & time resolution
NRLMSISE-00	SE	F10.7 | *ap* (array of 7 values)	30°, 3 h
JB2008	SE	S10, F10.7, M10, Y10 | *Dst*, *ap*	30°, 1 h
DTM2013	SE	F30 | *Kp*	30 × 15°, 3 h
TIE-GCM	FP	Solomon & Qian bands | Heelis-*Kp* Solomon & Qian bands | Weimer	5.0 × 5.0°, 1 min
CTIPe	FP	F10.7, F81, Solomon & Qian bands Weimer-2005, solar wind density and velocity, interplanetary magnetic field, Hemispheric Power, *Kp*	2.0 × 18°, 1 min

Eulerian continuity, momentum, and energy equations for the coupled thermosphere-ionosphere system (Roble et al., 1988; Richmond et al., 1992). It uses pressure surfaces as the vertical coordinate and extends in altitude from approximately 97 km to 600 km. The model resolution on the geographic grid employed throughout this study are 5° horizontal and 1/2 scale height H in the vertical, while a 2.5° horizontal and H/4 vertical resolution is also available. Tidal forcing at the lower boundary is specified by the Global Scale Wave Model (Hagan et al., 2001), and semi-annual and annual density periodicities are enhanced by applying seasonal variation of the eddy diffusivity coefficient at the lower boundary (Qian et al., 2014). Solar inputs are driven using $F_{10.7}$ radio solar flux measurements as a proxy for XUV/EUV/FUV solar flux as described by Solomon & Qian (2005), which were derived using the same model resolution and tidal lower boundary specification. The electrodynamo potential field is internally generated at middle and low latitudes using the model densities and neutral winds. This is merged with a magnetospheric potential at high latitudes, using one of two available empirically driven models. First, the Heelis et al. (1982) empirical formulation, driven by the *Kp* index, follows the method described in Solomon et al. (2012). Second, the Weimer empirical model, which uses upstream solar wind and IMF as input as described in the following subsections, is described by Weimer (2005). Separate simulations were carried out using each of these high-latitude potential models for this work. Recent developments include the addition of helium for high-altitude extension and lower boundary options as described in Qian et al. (2014), Sutton et al. (2015), and Maute (2017). The version 2.0 of TIE-GCM used in this work is a community release that was issued in March 2016.

2.3 CTIPe

The coupled thermosphere–ionosphere–plasmasphere electrodynamics (CTIPe) is a global, three-dimensional, time-dependent, nonlinear, self-consistent model that solves the momentum, energy, and composition equations for the neutral and ionized atmosphere (Fuller-Rowell et al., 1996; Millward et al. 2001; Codrescu et al., 2012). The global atmosphere in CTIPe is divided into a series of elements in geographic latitude, longitude, and pressure. The latitude resolution is 2°, the longitude resolution is 18°, and model parameters are calculated with a 1 min time step. In the vertical direction, the atmosphere is divided into 15 levels in logarithm of pressure from a lower boundary of 1 Pa at 80 km to more than 500 km altitude. The magnetospheric input is based on the statistical models of auroral precipitation and electric fields described by Fuller-Rowell & Evans (1987) and Weimer (2005), respectively. Auroral precipitation is keyed to the hemispheric power index (PI), based on the TIROS/NOAA auroral particle measurements. The Weimer electric field model is keyed to the solar wind parameters impinging the Earth's magnetosphere, and its input drivers include the magnitude of the interplanetary magnetic field (IMF) in the y–z plane, together with the velocity and density of the solar wind. A combination of measurements from Advanced Composition Explorer and Wind spacecraft instruments, obtained at NOAA Space Weather Prediction Center, NASA's Space Physics Data Facility and Los Alamos National Laboratory, have been used to address data gaps and quality issues (e.g., Skoug et al., 2004). The (2,2), (2,3), (2,4), (2,5), and (1,1) propagating tidal modes are imposed at 80 km altitude (Fuller-Rowell et al., 1991; Müller-Wodarg et al., 2001). The amplitudes and phases for the Hough modes are based on results from the Global Scale Wave Model-09 (GSWM-09; https://www2.hao.ucar.edu/gswm-global-scale-wave-model). In this paper, the lower boundary conditions in CTIPe simulations are specified using monthly averaged wind and temperature fields from the Whole Atmosphere Model (WAM; Akmaev et al., 2008; Fuller-Rowell et al., 2008). CTIPe uses time-dependent estimates of nitric oxide (NO) obtained from Marsh et al. (2004) empirical model based on student nitric oxide explorer (SNOE) satellite data rather than solving for minor species photochemistry self-consistently. For higher altitude applications, helium needs to be included in the model. Solar heating, ionization and dissociation rates, and their variation with solar activity are specified by Solomon & Qian (2005) solar EUV energy deposition scheme for upper atmospheric general circulation models.

3 Model assessment procedure

The next two subsections describe the density data and the necessary preprocessing, the phases and storms selected for comprehensive model evaluation by applying the new storm performance metrics. All models will be tested according to the same standards allowing unambiguous comparisons, and quantification of improvement of future model upgrades.

3.1 Selected density data

The density variability in the thermosphere is very large, typically hundreds of percent, on long time scales (the solar

Table 2. Datasets selected for the model assessment under storm conditions (*i* is inclination, and LST is the local solar time coverage and the approximate period, in days, to cover 24 h).

	Period	Altitude (km)	*i*	LST 24 h LST	Cadence	Precision (%)
CHAMP	05/2001–08/2010	450–250	87°	0–24 & 120–130	80 s	1–4%
GRACE	08/2002–07/2016	490–300	89°	0–24 & 120–160	80 s	2–6%
GOCE	11/2009–10/2013	270–180	96°	6–8 & 18–20	80 s	1–3%

Table 3. Selected storm intervals and type (* = coronal mass ejection; ** = high speed stream), minimum *Dst* and maximum *ap*/*Kp*, satellite density data (CH = CHAMP, GR = GRACE, GO = GOCE) and rounded local time at equator (hr).

Start & end date	Min *Dst* (nT)	Max *ap*/*Kp*	Density data (local time, hr)
19/Nov–22/Nov/2003*	−422	300/9−	CH (11/23), GR (3/15)
17/Jan–20/Jan/2005*	−103	179/8−	CH (8/20), GR (6/18)
21/Jan–24/Jan/2005*	−105	207/8	CH (8/20), GR (6/18)
07/May–10/May/2005**	−127	236/8+	CH (10/22), GR (11/23)
14/May–17/May/2005*	−263	236/8+	CH (10/22), GR (10/22)
29/May–02/Jun/2005*	−138	179/8−	CH (8/20), GR (9/21)
09/Jul–12/Jul/2005*	−92	94/6+	CH (4/16), GR (6/18)
23/Aug–26/Aug/2005*	−216	300/9−	CH (12/24), GR (2/14)
09/Sep–12/Sep/2005*	−147	179/8−	CH (10/22), GR (1/13)
14/Dec–17/Dec/2006*	−162	236/8+	CH (2/14), GR (6/18)
08/Mar–11/Mar/2012*	−131	207/8	GO (7/19)
16/Mar–19/Mar/2013*	−132	111/7−	GO (7/19)
31/May–03/Jun/2013**	−119	132/7	GR (7/19), GO (7/19)

cycle) but also on short time scales of hours to days in the event of strong geomagnetic storms. The variability depends on location, season, and the level of solar and geomagnetic activity, and it increases with altitude. However, most density measurements are in-situ along the satellite orbits, and the spatial and temporal data distribution of a single satellite is in fact rather poor. The latitudinal extent depends on the orbital inclination, and the local time coverage is essentially limited to the time of its ascending and descending pass, which changes slowly due to the precession of the orbital plane. The temporal resolution achieved with a single satellite is one orbital period of roughly 1.5 h (and then the satellite passes the same latitude and local time, but at a *different longitude*), even if the measurement cadence is 5 or 10 s. Precise density datasets inferred from accelerometer data of recent satellite missions are selected because only these are compatible with storm-time evaluation, notably precise measurements from pole to pole. However, the assessments are based on densities from one, or two satellites at best, and one cannot reconstruct the complete picture of the thermosphere at any given time; we only see the densities in a latitude-local time frame at a relatively constant altitude, with a temporal resolution of about 95 min.

Table 2 lists the essential information of the three selected datasets, CHAMP (Doornbos, 2011), GRACE (Bruinsma; unpublished) and GOCE (Bruinsma et al., 2014). The densities used in this study were smoothed to suppress variations with scales smaller than 600 km, and then down-sampled to 80 s cadence. The original GOCE densities can be obtained, after registration, on the ESA server (https://earth.esa.int).

Table 3 lists the dates, minimum *Dst* and maximum *ap*/*Kp* of the storms, and the satellite data available for model assessment. The intervals are selected to cover a strong storm sequence that returns to low geomagnetic activity (low *Kp*), and secondly, observations from two satellites are preferred. This led to a slightly different selection of storms than proposed in (Bruinsma et al., 2018). Thirteen storms, eleven of which were CME-driven and two by High Speed Streams, were selected for this assessment using the available data provided by the three satellites. The eight storms in 2005 are the so-called problem storms for the US Air Force (Knipp et al., 2013). Strong storms cause the largest satellite orbit perturbations, together with degraded tracking performance, and model assessment is most pertinent and needed for those events. The impact of weak storms is not dramatic, and only one is selected because

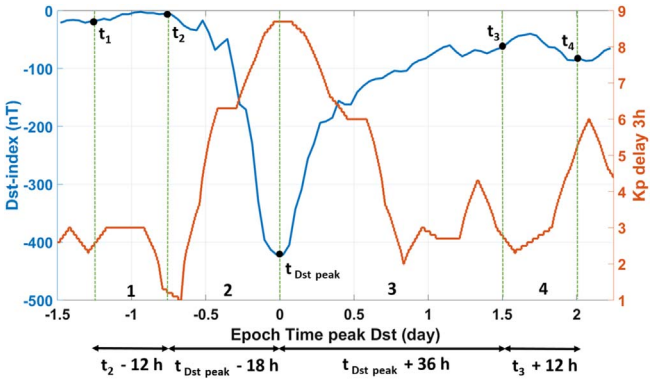

Fig. 1. The four phases of the assessment interval, centered on the time of minimum of (hourly) *Dst*. The 3-hourly *Kp* index is also shown because it used in DTM2013, and NRLMSISE-00 after conversion to *ap*.

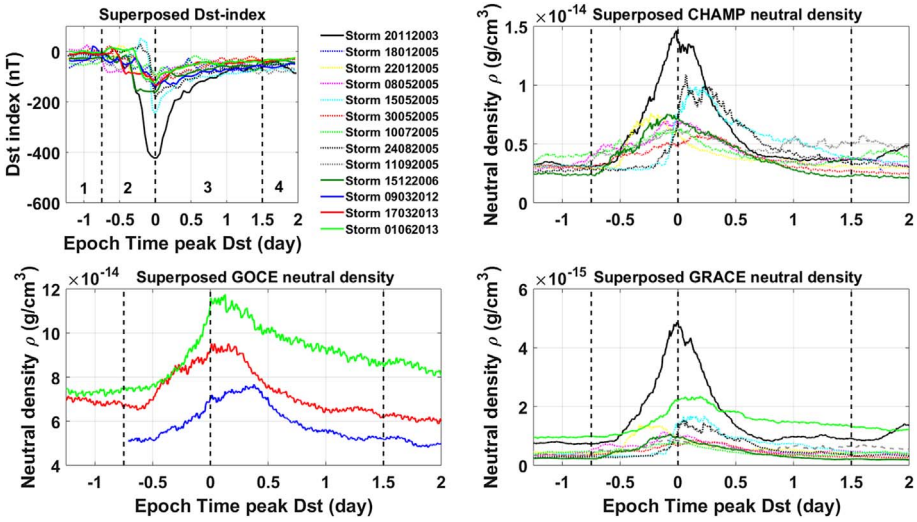

Fig. 2. *Dst* (top) and densities, smoothed over one orbit in a moving window, for the selected storms, centered on t_0, for CHAMP, GRACE and GOCE.

it was a problem storm. While not strong but of moderate strength, the three storms in 2012 and 2013 were selected because they are the only ones for which high resolution density data is available in solar cycle 24, from GRACE and GOCE. The local time at the equator is also given in Table 3, and it is clear that the distribution is sparse for any single storm.

3.2 Metrics for model-data comparison

The updated metrics for storms differ from (Bruinsma et al., 2018) mainly in how the time interval for assessment is defined. Storms are divided in four phases, two before and two after the minimum *Dst* value. After verifying that differences with a physics-based definition of the phases are small, and to facilitate automation, it was decided to use intervals of fixed lengths for the phases. The phases correspond to pre-storm (1), onset (2), recovery (3), and post-storm (4). Figure 1 illustrates these four phases with respect to the minimum in *Dst*. The pre-storm interval is used to de-bias the model with respect to the observations by computing a scaling factor, which is then applied to the model densities in phases 1–4. The scaling factor is determined by computing the ratio of the sum of all observations to the sum of all model densities in phase 1. This de-biasing procedure is used to minimize the effect of non-storm related model errors on the assessment. All data for each storm are selected 30-h before to 48-h after the minimum in *Dst*, which is defined as t_0. The *Dst* and densities centered on t_0 for all storms listed in Table 3 are displayed per satellite in Figure 2. The choice of fixed lengths of the intervals is supported by the density profiles shown in Figure 2. Note that GOCE is missing data in phase 1 for the storm in 2012; there is no data for GRACE. This example of a data gap in one of the phases, maintained in this analysis because of the few observed storms in solar cycle 24, is another reason for the sparseness of the storm density database. Data is regularly missing during storms. In this case for GOCE in 2012, the scaling factor for de-biasing was not determined according to the metrics, but with part of the data in phase 2.

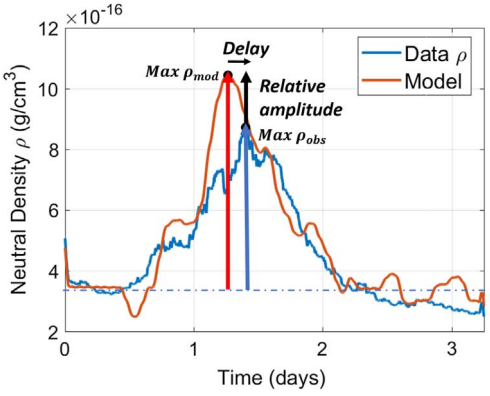

Fig. 3. Metrics of the storm peak amplitude and timing assessment. The model density is de-biased.

The models and data can be compared by computing density residuals, which is an absolute difference (observed minus computed). A better quantity to express a model's skill to reproduce the observations, i.e. reality, is the observed-to-computed (O/C) density ratio. Density ratios of one indicate perfect duplication of the observations, i.e. an unbiased model that reproduces all features; deviation from unity points to under (larger than one) or overestimation (smaller than one). Because of the very large and dynamic range in density, mainly due to differences in altitude and solar activity (i.e. phase of the solar cycle), it is rather difficult to analyze and interpret model performance in absolute values. The relative precision given in the form of density ratios is always simple to comprehend. A model bias, i.e. the mean of the density ratios differs from unity, is most damaging to orbit extrapolation because it causes position errors that increase with time. The standard deviation (SD) of the density ratios, computed as percentage of the observation, represents a combination of the ability of the model to reproduce observed density variations, and the geophysical noise (e.g. waves, the short duration effect of large flares) and

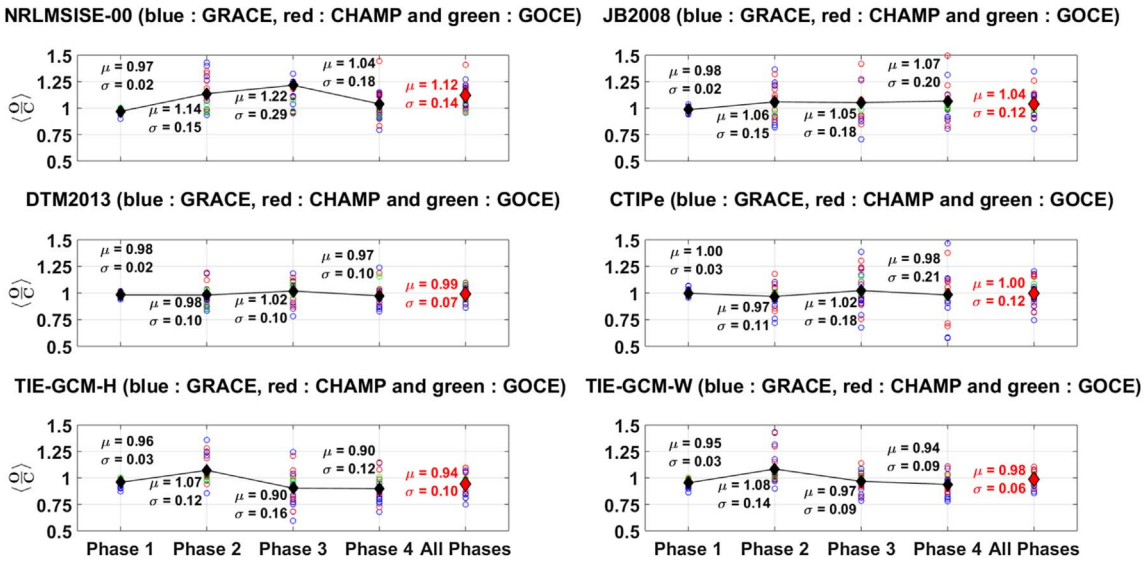

Fig. 4. The mean μ and the standard deviation σ of the 24 mean density ratios, per phase (black) and overall (red), using data from three satellites.

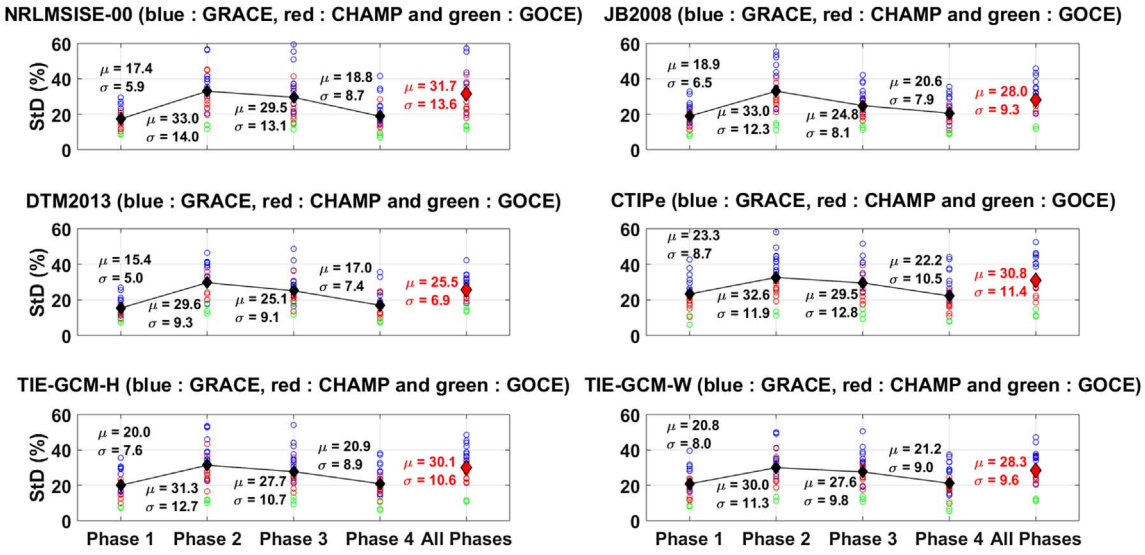

Fig. 5. The mean μ and the standard deviation σ of the 24 standard deviations of the density ratios (SD; %), per phase (black) and overall (red), using data from three satellites.

instrumental noise in the observations. The mean and SD of the density ratios, due to their distribution, are computed in Log space (Sutton, 2018):

$$\text{Mean}\left(\frac{O}{C}\right) = \exp\left(\frac{1}{N}\sum_{n=1}^{N}\ln\frac{O_n}{C_n}\right);\ \text{SD}\left(\frac{O}{C}\right) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\ln\frac{O_n}{C_n} - \ln\text{Mean}\left(\frac{O}{C}\right)\right)^2} \quad (1)$$

where N is the total number of observations. The correlation coefficients R are also computed. The correlation coefficient is independent of model bias and R^2 represents the fraction of observed variance captured by the model. Mean, SD and correlation are computed for each storm, for each separate phase as well as for the entire interval. These metrics are the same as in (Bruinsma et al., 2018) but applied to the defined storm intervals only in order to isolate the performance of the geomagnetic storm algorithm of the models.

A second assessment, and updated metrics compared with (Bruinsma et al., 2018), concerns the amplitude and timing of the maximum density peak considering the entire time interval. The absolute relative amplitude error is expressed as a percentage of the measured maximum, and the timing of the peak with respect to the observed peak is expressed in hours, an example of which is shown in Figure 3. However, these two quantities cannot always be determined unambiguously, for example when two peaks are present, or a broad peak is present. For that

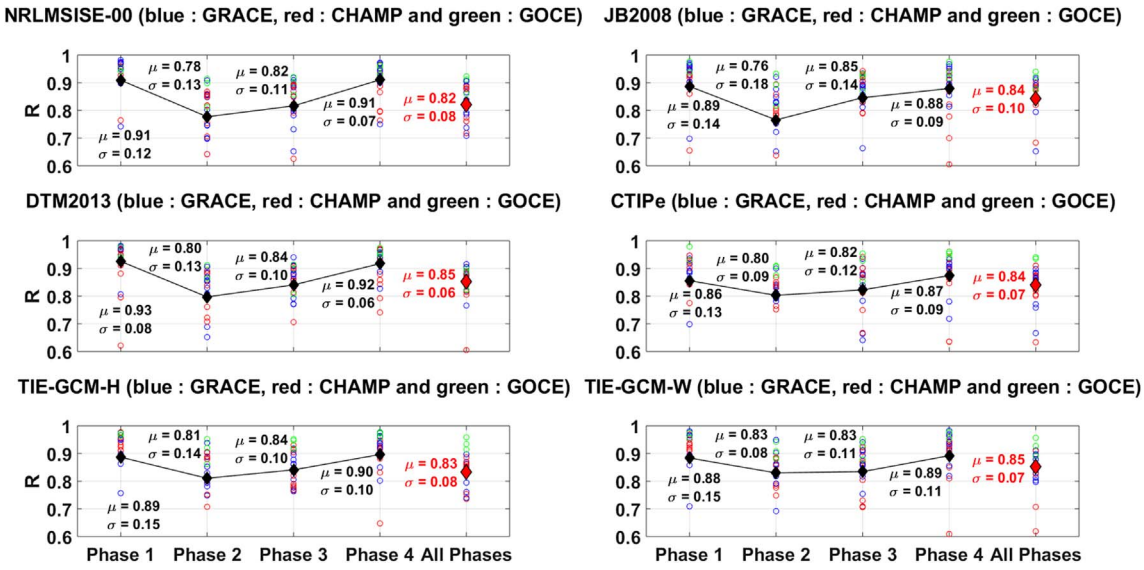

Fig. 6. The mean μ and standard deviation σ of the 24 correlation coefficients per phase (black) and overall (red), using data from three satellites.

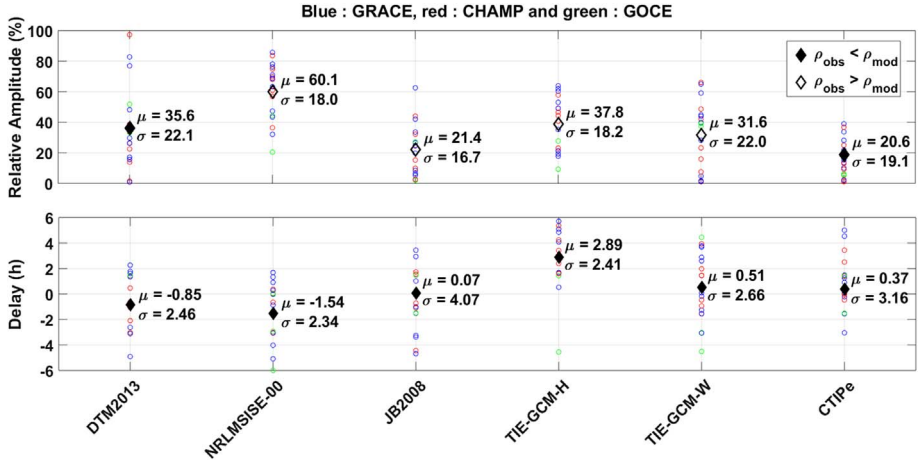

Fig. 7. The absolute relative amplitude difference (top) and the modeled minus observed delay in time with respect to the maximum density peak, i.e. negative values mean that the model peaks too early (bottom) for 11 storms, using data from three satellites.

reason, we have in the next section rejected results that could not be well determined.

4 Storm-time assessment results

The means of the density ratios per phase are displayed in Figure 4 for the five models, using different colors per satellite. The printed numbers are the means (μ) and standard deviations (σ) of the 24 colored symbols plotted vertically for each phase (in black) and for the entire storm event (phases 1–4; in red). The de-biasing (i.e. applying the scaling factor determined in phase 1 to all model densities in phases 1–4) results in density ratios close to unity in phase 1, and less scatter (smaller σ) in all phases. The TIE-GCM runs with the Heelis and Weimer models as drivers are named TIEGCM-H and TIEGCM-W, respectively. From the satellite operational point of view, the overall mean is the most important number to consider as it directly relates to the satellite position at the end of the storm. It informs on the performance of the model over a complete storm, and a mean density ratio of one means that thermospheric density (proportional to satellite aerodynamic drag) was correct on average. Densities that were predicted too large compensated those too low and vice versa during the storm interval, leading to a correct mean density and consequently a correct satellite position at the end of the storm (but not necessarily during the storm). The most accomplished SE (FP) model according to this criterion is DTM2013 (TIEGCM-W), which obtains a mean of 0.99, i.e. only 1% mean bias, and a small standard deviation of 0.07. DTM2013, JB2008 and CTIPe have stable means per phase, i.e. the bias does not evolve significantly as a function of storm activity. NRLMSISE-00 has a clear storm signature, underestimating the density during the storm phases 2 and even more during phase 3. TIEGCM underestimates density

in Phase 2, but the Weimer model radically improves performance in the decaying storm phase (Phase 3), enhancing the mean from 0.90 to 0.97 and reducing the standard deviation from 0.16 to 0.09.

The mean standard deviations of the density ratios (SD; %) per phase are displayed in Figure 5. The mean standard deviations for GOCE (green symbols) is smallest, followed by CHAMP (red symbols), and GRACE (blue symbols) has the largest mean standard deviation; this nicely demonstrates that relative variability increases with altitude, although it is partly due to data becoming noisier too. All models display significant degradation during the main storm phases 2 and 3, and all have the worst performance (largest standard deviation) for phase 2. The standard deviation of the SE models nearly doubles from phase 1 to 2, and the largest standard deviations are seen for NRLMSISE-00 and JB2008 (33.0%). JB2008 then achieves the best performance for phase 3, while NRLMSISE-00 and CTIPe have the largest mean standard deviations (29.5%) for that phase. The most accomplished SE and FP models over the entire storm are DTM2013 and TIEGCM-W, which have means of 25.5% and 28.3%, respectively. However, the positive impact on the standard deviation thanks to using the Weimer model instead of the Heelis model in TIE-GCM is very modest.

The mean correlation coefficients per phase are displayed in Figure 6. As expected, all models have reduced correlations in phases 2 and 3, and the lowest correlations are reached in phase 2. In line with results shown in Figures 4 and 5, the highest mean correlations over the entire storm are attained with DTM2013 and TIEGCM-W, which have a mean of 0.85.

The amplitudes and phases of the peaks in density are compared for 11 storms (18/01/2005 and 11/09/2005 do not allow unambiguous testing) and the results are displayed in Figure 7. The amplitudes are on average underestimated with NRLMSISE-00, JB2008, and TIE-GCM, whereas DTM2013 and CTIPe overestimated. CTIPe and JB2008 estimate the amplitudes with the smallest mean difference of slightly over 20%, while NRLMSISE-00 underestimates on average by 60%. DTM2013 and NRLMSISE-00 on average predict the storm peaks too early, while TIE-GCM and CTIPe by a small amount (0.37 h) predict them too late. The phasing of the storm peak is best on average with JB2008, which has a 4-min mean delay; however, the standard deviation is largest of all models, 4 h. The largest mean timing error of 2.89 h is reached with TIEGCM-H; using Weimer instead of the Heelis model improves the results considerably to an average delay of 0.51 h (amplitude is more correct too).

5 Summary and conclusions

The density data and indices for the 13 selected storms as well as updated metrics for thermosphere model assessment have been described in this paper. All storms are divided in four phases, which are relative to the time of minimum *Dst*. The mean and standard deviation of the density ratios, the correlations and the amplitude and timing of the peak density, are computed using available CHAMP, GRACE and GOCE density data and the CIRA models NRLMSISE-00, JB2008 and DTM2013, and the first principles models CTIPe and TIE-GCM using Heelis or Weimer drivers. The best results over the entire 4-phase storm period are obtained with DTM2013 and TIEGCM-W, while the oldest model, NRLMSISE-00, is the least precise. Compared to the assessments presented in (Bruinsma et al., 2018), in which the same models except TIEGCM-W were evaluated using entire years of data, this study confirms that best results are obtained with DTM2013 and that NRLMSISE-00 is trailing. During storms, TIEGCM-W is more precise than JB2008, NRLMSISE-00 and CTIPe, and using the Weimer instead of the Heelis model has a large impact during storms. The standard deviation of the density ratios increases with altitude, as in (Bruinsma et al., 2018), but around 450 km (GRACE) they are 40–60% during storms, which is 2–3 times larger than when comparing over entire years.

An important result is that the means of the density ratios of DTM2013, JB2008 and CTIPe do not significantly depend on the phase 1–4 of the storm, even if the standard deviations very much do for all models. NRLMSISE-00 underestimates during the onset and decaying phases (2 and 3) of storms, TIEGCM-H underestimates phase 2 and overestimates in phase 3, and TIEGCM-W underestimates phase 2. The results per phase, and the fidelity of the storm peak amplitude and time of maximum, showed strengths and weaknesses that the model developers can presently focus on.

The current model assessment is far from comprehensive. A more thorough model assessment, e.g. solar cycle, local time and altitude effects, under storm conditions requires more and better-distributed observations of density. In this study, with data from two (or one) satellites, density variations are monitored in four local time planes at best, and that only at two altitudes. Only four $Kp = 9$ storms were observed since 2000, thanks to CHAMP, GRACE and GOCE; bear in mind that the assessment is possible only thanks to *data of opportunity*, the objective of those missions was geodesy.

Acknowledgements. We thank ESA (GOCE densities at earth.esa.int), NASA (GRACE), and DLR (CHAMP and GRACE) for making their data publicly available. We thank WDC Kyoto for the *Dst* index (http://wdc.kugi.kyoto-u.ac.jp/), and GFZ Potsdam for the *Kp* index (https://www.gfz-potsdam.de/en/kp-index/). SB and CB received support from CNES/SHM and the SWAMI project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776287. MF was supported by NSF grant AGS1651459 and NASA grant NNX17AI34G. ES was supported by the University of Colorado Chancellor's Office Grand Challenge Initiative "Our Space. Our Future" via the Space Weather Technology, Research, and Education Center. We acknowledge the reviewers for their valuable comments that helped to improve the manuscript. The editor thanks Astrid Maute and an anonymous reviewer for their assistance in evaluating this paper.

References

Akmaev RA, Fuller-Rowell TJ, Wu F, Forbes JM, Zhang X, Anghel AF, Iredell MD, Moorthi S, Juang HM. 2008. Tidal variability in the lower thermosphere: Comparison of whole atmosphere model (WAM) simulations with observations from TIMED. *Geophys Res Lett* **35**: L03810. https://doi.org/10.1029/2007GL032584.

Bowman B, Tobiska WK, Marcos F, Huang C, Lin C, Burke W. 2008. A new empirical thermospheric density model JB2008 using new solar and geomagnetic indices. In: *AIAA/AAS Astrodynamics Specialist Conference*, AIAA, Honolulu, HI.

Bruinsma SL. 2015. The DTM-2013 thermosphere model. *J Space Weather Space Clim* **5**: A1. https://doi.org/10.1051/swsc/2012005.

Bruinsma SL, Forbes JM. 2007. Global observation of traveling atmospheric disturbances (TADs) in the thermosphere. *Geophys Res Lett* **34**: L14103. https://doi.org/10.1029/2007GL030243.

Bruinsma S, Forbes JM, Nerem S, Zhang X. 2006. Thermosphere density response to the 20–21 November 2003 solar and geomagnetic storm from CHAMP and GRACE accelerometer data. *J Geophys Res* **11**: A06303. https://doi.org/10.1029/2005JA011284.

Bruinsma SL, Doornbos E, Bowman BR. 2014. Validation of GOCE densities and thermosphere model evaluation. *Adv Space Res* **54**: 576–585. https://doi.org/10.1016/j.asr.2014.04.008.

Bruinsma S, Sutton E, Solomon SC, Fuller-Rowell T, Fedrizzi M. 2018. Space weather modeling capabilities assessment: Neutral density and orbit determination at LEO. *Space Weather* **16(11)**: 1806–1816. https://doi.org/10.1029/2018SW002027.

Bussy-Virat CD, Ridley AJ, Getchius JW. 2018. Effects of uncertainties in the atmospheric density on the probability of collision between space objects. *Space Weather* **16**: 519–537. https://doi.org/10.1029/2017SW001705.

Champion KSW, Marcos FA. 1973. The triaxial-accelerometer system on atmosphere explorer. *Radio Sci* **8**: 297–303.

Codrescu MV, Negrea C, Fedrizzi M, Fuller-Rowell TJ, Dobin A, Jakowsky N, Khalsa H, Matsuo T, Maruyama N. 2012. A real-time run of the Coupled Thermosphere Ionosphere Plasmasphere Electrodynamics (CTIPe) model. *Space Weather* **10**: S02001. https://doi.org/10.1029/2011SW000736.

Doornbos E. 2011. Thermospheric density and wind determination from satellite dynamics. *PhD Dissertation*, University of Delft, Delft, the Netherlands, 188 p. Available at http://repository.tudelft.nl/.

Emmert JT. 2015. Altitude and solar activity dependence of 1967–2005 thermospheric density trends derived from orbital drag. *J Geophys Res* **120**: 2940–2950. https://doi.org/10.1002/2015JA021047.

Forbes JM, Roble RG, Marcos FA. 1987. Thermospheric dynamics during the March 22, 1979 magnetic storm: 2. Comparisons of model predictions with observations. *J Geophys Res* **92**: 6069–6081.

Forbes JM, Lu G, Bruinsma S, Nerem S, Zhang X. 2005. Thermosphere density variations due to the 15–24 April 2002 solar events from CHAMP/STAR accelerometer measurements. *J. Geophys. Res.* **110**: A12S27. https://doi.org/10.1029/2004JA010856.

Fuller-Rowell TJ, Evans DS. 1987. Height integrated Pedersen and Hall conductivity patterns inferred from the TIROS-NOAA satellite data. *J Geophys Res* **92**: 7606–7618.

Fuller-Rowell TJ, Rees D, Quegan S, Moffett RJ. 1991. Numerical simulations of the sub-auroral F-region trough. *J Atmos Terr Phys* **53**: 529–540.

Fuller-Rowell TJ, Rees D, Quegan S, Moffett RJ, Codrescu MV, Millward GH. 1996. A coupled thermosphere-ionosphere model (CTIM). In: *Handbook of ionospheric models*, Schunk RW (Ed.), Utah State Univ, Logan, Utah, pp. 217–238.

Fuller-Rowell TJ, Akmaev RA, Wu F, Anghel A, Maruyama N, Anderson DN, Codrescu MV, Iredell M, Moorthi S, Juang HM, Hou YT, Millward G. 2008. Impact of terrestrial weather on the upper atmosphere. *Geophys Res Lett* **35**: L09808. https://doi.org/10.1029/2007GL032911.

Hagan ME, Roble RG, Hackney J. 2001. Migrating thermospheric tides. *J Geophys Res* **106**: 12739–12752. https://doi.org/10.1029/2000JA000344.

Heelis RA, Lowell JK, Spiro RW. 1982. A model of the high-latitude ionosphere convection pattern. *J Geophys Res* **87**: 6339. https://doi.org/10.1029/JA087iA08p06339.

Hejduk MD, Snow DE. 2018. The effect of neutral density estimation errors on satellite conjunction serious event rates. *Space Weather* **16**: 849–869. https://doi.org/10.1029/2017SW001720.

Jacchia LG, Slowey J. 1963. Accurate drag determinations for eight artificial satellites: Atmospheric densities and temperatures. *Smithson Contrib Astrophys* **8**: 1–99.

Kalafatoglu Eyiguler EC, Shim JS, Kuznetsova MM, Kaymaz Z, Bowman BR, et al. 2019. Quantifying the storm time thermospheric neutral density variations using model and observations. *Space Weather* **17**: 269–284. https://doi.org/10.1029/2018SW002033.

Knipp D, Kilcommons L, Hunt L, Mlynczak M, Pilipenko V, Bowman B, Deng Y, Drake K. 2013. Thermospheric damping response to sheath-enhanced geospace storms. *Geophys Res Lett* **40**: 1263–1267. https://doi.org/10.1002/grl.50197.

Knipp DJ, Pette DV, Kilcommons LM, Isaacs TL, Cruz AA, Mlynczak MG, Hunt LA, Lin CY. 2017. Thermospheric nitric oxide response to shock-led storms. *Space Weather* **15**: 325–342. https://doi.org/10.1002/2016SW001567.

Liu H, Luehr H. 2005. Strong disturbance of the upper thermospheric density due to magnetic storms: CHAMP observations. *J Geophys Res* **110**: A09S29. https://doi.org/10.1029/2004JA010908.

Marsh DR, Solomon SC, Reynolds AE. 2004. Empirical model of nitric oxide in the lower thermosphere. *J Geophys Res* **109**: A07301. https://doi.org/10.1029/2003JA010199.

Maute A. 2017. Thermosphere-ionosphere-electrodynamics general circulation model for the ionospheric connection explorer: TIEGCM-ICON. *Space Sci Rev* **212**: 523–551. https://doi.org/10.1007/s11214-017-0330-3.

Mehta PM, Walker AC, Sutton EK, Godinez HC. 2017. New density estimates derived using accelerometers on board the CHAMP and GRACE satellites. *Space Weather* **15**: 558–576. https://doi.org/10.1002/2016SW001562.

Millward GH, Müller-Wodarg ICF, Aylward AD, Fuller-Rowell TJ, Richmond AD, Moffett RJ. 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. *J Geophys Res* **106**: 24733–24744.

Müller-Wodarg ICF, Aylward AD, Fuller-Rowell TJ. 2001. Tidal oscillations in the thermosphere: A theoretical investigation of their sources. *J Atmos Sol-Terr Phys* **63**: 899–914.

Nier AO, Potter WE, Hickman DR, Mauersberger K. 1973. The open-source neutral-mass spectrometer on atmosphere explorer-C, -D, -E. *Radio Sci* **8**: 271–276.

Pedatella NM, Fuller-Rowell T, Wang H, Jin H, Miyoshi Y, et al. 2014. The neutral dynamics during the 2009 sudden stratosphere warming simulated by different whole atmosphere models. *J Geophys Res* **119**: 1306–1324. https://doi.org/10.1002/2013JA019421.

Picone JM, Hedin AE, Drob DP, Aikin AC. 2002. NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *J Geophys Res* **107(A12)**: 1468. https://doi.org/10.1029/2002JA009430.

Qian L, Burns AG, Emery BA, Foster B, Lu G, Maute A, Richmond AD, Roble RG, Solomon SC, Wang W. 2014. The NCAR TIE-GCM: A community model of the coupled thermosphere/ionosphere system. In: *Modeling the ionosphere-thermosphere system*, Huba J, Schunk R, Khazanov G (Eds.),AGU Geophysical Monograph Series, Vol. **201**, pp. 73. https://doi.org/10.1002/9781118704417.ch7.

Richmond AD, Ridley EC, Roble RG. 1992. A thermosphere/ionosphere general circulation model with coupled electrodynamics. *Geophys Res Lett* **19**: 601. https://doi.org/10.1029/92GL00401.

Roble RG, Ridley EC, Richmond AD, Dickinson RE. 1988. A coupled thermosphere/ionosphere general circulation model. *Geophys Res Lett* **15**: 1325. https://doi.org/10.1029/GL015i012p01325.

Skoug RM, Gosling JT, Steinberg JT, McComas DJ, Smith CW, Ness NF, Hu Q, Burlaga LF. 2004. Extremely high speed solar wind: 29–30 October 2003. *J Geophys Res* **109**: A09102. https://doi.org/10.1029/2004JA010494.

Solomon SC, Qian L. 2005. Solar extreme-ultraviolet irradiance for general circulation models. *J Geophys Res* **110**: A10306. https://doi.org/10.1029/2005JA011160.

Solomon SC, Burns AG, Emery BA, Mlynczak MG, Qian L, Wang W, Weimer DR, Wiltberger M. 2012. Modeling studies of the impact of high-speed streams and co-rotating interaction regions on the thermosphere-ionosphere. *J Geophys Res* **117**: A00L11. https://doi.org/10.1029/2011JA017417.

Storz MF, Bowman BR, Branson MJI, Casali SJ, Tobiska WK. 2005. High accuracy satellite drag model (HASDM). *Adv Space Res* **36**: 2497–2505. https://doi.org/10.1016/j.asr.2004.02.020.

Sutton EK. 2018. A new method of physics-based data assimilation for the quiet and disturbed thermosphere. *Space Weather* **16**: 736–753. https://doi.org/10.1002/2017SW00178.

Sutton EK, Thayer JP, Wang W, Solomon SC, Liu X, Foster BT. 2015. A self-consistent model of helium in the thermosphere. *J Geophys Res* **120**: 6884–6900. https://doi.org/10.1002/2015JA021223.

Weimer DR. 2005. Improved ionospheric electrodynamic models and application to calculating Joule heating rates. *J Geophys Res* **110**: A05306. https://doi.org/10.1029/2004JA010884.

Cite this article as: Bruinsma S, Boniface C, Sutton EK & Fedrizzi M 2021. Thermosphere modeling capabilities assessment: geomagnetic storms. *J. Space Weather Space Clim.* **11**, 12. https://doi.org/10.1051/swsc/2021002.