Validation of Ionospheric Specifications During Geomagnetic Storms: TEC and foF2 during the 2013 March Storm Event-II

Ja Soon Shim[1], In-Sun Song[2], Geonhwa Jee[3], Young-Sil Kwak[4], Ioanna Tsagouri[5], Larisa Goncharenko[6], Joseph McInerney[7], Francis Vitt[7], Lutz Rastaetter[8], Jia Yue[9], Min-Yang Chou[9], Mihail V. Codrescu[10], Anthea J Coster[11], Mariangel Fedrizzi[12], Timothy J. Fuller-Rowell[13], Aaron J. Ridley[14], and Stanley C. Solomon[7]

[1]Yonsei University,
[2]Yonsei University
[3]Korea Polar Research Institute
[4]Korea Astronomy and Space Science Institute
[5]National Observatory of Athens
[6]MIT Haystack Observatory, Westford, MA, USA.
[7]National Center for Atmospheric Research (UCAR)
[8]NASA/GSFC
[9]Goddard Space Flight Center
[10]Space Weather Prediction Center, National Oceanic And Atmospheric Administration
[11]MIT Haystack Observatory
[12]University of Colorado/CIRES and NOAA/SWPC
[13]NOAA Space Weather Prediction Center
[14]University of Michigan-Ann Arbor

December 18, 2022

Abstract

Assessing space weather modeling capability is a key element in improving existing models and developing new ones. In order to track improvement of the models and investigate impacts of forcing, from the lower atmosphere below and from the magnetosphere above, on the performance of ionosphere-thermosphere models, we expand our previous assessment for 2013 March storm event [Shim et al., 2018]. In this study, we evaluate new simulations from upgraded models (Coupled Thermosphere Ionosphere Plasmasphere Electrodynamics (CTIPe) model version 4.1 and Global Ionosphere Thermosphere Model (GITM) version 21.11) and from NCAR Whole Atmosphere Community Climate Model with thermosphere and ionosphere extension (WACCM-X) version 2.2 including 8 simulations in the previous study. A simulation of NCAR Thermosphere-Ionosphere-Electrodynamics General Circulation Model version 2 (TIE-GCM 2) is also included for comparison with WACCM-X. TEC and foF2 changes from quiet-time background are considered to evaluate the model performance on the storm impacts. For evaluation, we employ 4 skill scores: Correlation coefficient (CC), root-mean square error (RMSE), ratio of the modeled to observed maximum percentage changes (Yield), and timing error(TE). It is found that the models tend to underestimate the storm-time enhancements of foF2 (F2-layer critical frequency) and TEC (Total Electron Content) and to predict foF2 and/or TEC better in the North America but worse in the Southern Hemisphere. The ensemble simulation for TEC is comparable to results from a data assimilation model (Utah State University-Global Assimilation of Ionospheric Measurement (USU-GAIM)) with differences in skill score less than 3% and 6% for CC and RMSE, respectively.

Validation of Ionospheric Specifications During Geomagnetic Storms: TEC and foF2 during the 2013 March Storm Event-II

J. S. Shim[1], I.-S. Song[1], G. Jee[2], Y.-S. Kwak[3], I. Tsagouri[4], L. Goncharenko[5], J. McInerney[6], A. Vitt[6], L. Rastaetter[7], J. Yue[7,8], M. Chou[7,8], M. Codrescu[9], A. J. Coster[5], M. Fedrizzi[9], T. J. Fuller-Rowell[9], A. J. Ridley[10], S. C. Solomon[6]

[1]Department of Atmospheric Sciences, Yonsei University, Seoul, South Korea,

[2]Division of Atmospheric Sciences, Korea Polar Research Institute, Incheon, South Korea

[3]Space Science Division, Korea Astronomy and Space Science Institute, Daejeon, South Korea

[4]National Observatory of Athens, Penteli, Greece,

[5]Haystack Observatory, Westford, MA, USA,

[6]High Altitude Observatory, NCAR, Boulder, CO, USA,

[7]NASA GSFC, Greenbelt, MD, USA,

[8]Catholic University of America, Washington, DC, USA,

[9]NOAA SWPC, Boulder, CO, USA,

[10]Space Physics Research Laboratory, Univ. of Michigan, Ann Arbor, MI, USA

Corresponding author: In-Sun Song (songi@yonsei.ac.kr)

Key Points:

- foF2/TEC and their changes during a storm predicted by seven ionosphere-thermosphere coupled models are evaluated against GIRO foF2 and GPS TEC measurements.
- Model simulations tend to underestimate the storm-time enhancements of foF2 and TEC and to predict them better in the North America but worse in the southern hemisphere.
- Ensemble of all simulations for TEC is comparable to the data assimilation model (USU-GAIM).

Abstract

Assessing space weather modeling capability is a key element in improving existing models and developing new ones. In order to track improvement of the models and investigate impacts of forcing, from the lower atmosphere below and from the magnetosphere above, on the performance of ionosphere-thermosphere models, we expand our previous assessment for 2013 March storm event [*Shim et al.*, 2018]. In this study, we evaluate new simulations from upgraded models (Coupled Thermosphere Ionosphere Plasmasphere Electrodynamics (CTIPe) model version 4.1 and Global Ionosphere Thermosphere Model (GITM) version 21.11) and from NCAR Whole Atmosphere Community Climate Model with thermosphere and ionosphere extension (WACCM-X) version 2.2 including 8 simulations in the previous study. A simulation of NCAR Thermosphere-Ionosphere-Electrodynamics General Circulation Model version 2 (TIE-GCM 2) is also included for comparison with WACCM-X. TEC and foF2 changes from quiet-time background are considered to evaluate the model performance on the storm impacts. For evaluation, we employ 4 skill scores: Correlation coefficient (CC), root-mean square error (RMSE), ratio of the modeled to observed maximum percentage changes (Yield), and timing error(TE). It is found that the models tend to underestimate the storm-time enhancements of foF2 (F2-layer critical frequency) and TEC (Total Electron Content) and to predict foF2 and/or TEC better in the North America but worse in the Southern Hemisphere. The ensemble simulation for TEC is comparable to results from a data assimilation model (Utah State University-Global Assimilation of Ionospheric Measurement (USU-GAIM)) with differences in skill score less than 3% and 6% for CC and RMSE, respectively.

Plain Language Summary

The Earth's ionosphere-thermosphere (IT) system, which is present between the lower atmosphere and the magnetosphere, is highly variable due to external forcings from below and above as well as internal forcings mainly associated with ion-neutral coupling processes. The variabilities of the IT system can adversely affect our daily lives, therefore, there is a need for both accurate and reliable weather forecasts to mitigate harmful effects of space weather events. In order to track the improvement of predictive capabilities of space weather models for the IT system, and to investigate the impacts of the forcings on the performance of IT models, we evaluate new simulations from upgraded models (CTIPe model version 4.1 and GITM version 21.11) and from NCAR WACCM-X version 2.2 together with 8 simulations in the previous study. A simulation of NCAR TIE-GCM version 2 is also included for the comparison with WACCM-X. Quantitative evaluation is performed by using 4 skill scores including Correlation coefficient (CC), root-mean square error (RMSE), ratio of the modeled to observed maximum percentage changes (Yield), and timing error (TE). The findings of this study will provide a baseline for future validation studies of new and improved models.

1. Introduction

Variabilities of the Earth's ionosphere-thermosphere (IT) system, caused by charged particles and electromagnetic radiation emitted from the sun, can adversely affect our daily lives, which are highly dependent on space-based technological infrastructures such as Low-Earth Orbit (LEO) satellites and the Global Navigation Satellite System (GNSS). To mitigate harmful effects of space weather events, modeling plays a critical role in our quest to understand the connection between solar eruptive phenomena and their impacts in interplanetary space and near-Earth space environment. In particular, the Earth's upper atmosphere including the IT system is the space environment closest to the human society. Thus, during the past few decades, first-principles physics-based (PB) IT models have been developed for specifications and forecasts of the near-Earth space environment. In addition, there have been recent developments of whole atmosphere models with thermospheric and ionospheric extension to fully understand variabilities of the IT system by considering coupling between the IT system and the lower atmosphere [e.g., *Akmaev,* 2011; *Fuller-Rowell et al.,* 2010; *Jin et al*., 2011; *Liu et al*., 2018].

For more accurate space weather forecasting, assessing space weather modeling capability is a key element to improve existing models and to develop new models. Over the last decade, in an effort to address the needs and challenges of the assessment of our current knowledge about space weather effects on the IT system and current state of IT modeling capabilities, the NASA GSFC Community Coordinated Modeling Center (CCMC) has been supporting community-wide model validation projects, including Coupling, Energetics and Dynamics of Atmospheric Regions (CEDAR) [*Shim et al.*, 2011, 2012, 2014] and Geospace Environment Modeling (GEM)-CEDAR modeling challenges [*Rastätter et al*., 2016; *Shim et al.*, 2017a].

Furthermore, in 2018, the CCMC established an international effort, “International Forum for Space Weather Modeling Capabilities Assessment”, to evaluate and assess the predictive capabilities of space weather models (https://ccmc.gsfc.nasa.gov/iswat/IFSWCA/). As a result of this international effort, four ionosphere/thermosphere working groups were established with an overarching goal to devise a standardized quantitative validation procedure for IT models [*Scherliess et al*., 2019].

The working group, focusing on neutral density and orbit determination at LEO, reported their initial results for specific metrics for thermosphere model assessment over the selected three full years and two geomagnetic storms in 2005 [*Bruinsma et al.*, 2018]. They reported that the tested models in general performed reasonably well, although seasonal errors were sometimes observed and impulsive geomagnetic events remain a challenge. Kalafatoglu Eyigüler et al. (2019) compared the neutral density estimates from two empirical and three PB models with those obtained from the CHAMP satellite. They suggested that several metrics that provide different aspects of the errors should be considered together for a proper performance evaluation.

Another working group, "Ionosphere Plasmasphere Density Working Team", performed the assessment of present modeling capabilities in predicting the ionospheric climatology of f_oF_2 and hmF2 for the entire year 2012 [*Tsagouri et al.,* 2018]. Tsagouri et al. (2018) identified a strong seasonal and local time dependence of the model performances, especially for PB models, which could provide useful insight for future model improvements. Tsagouri et al. cautioned that the quality of the ground truth data may play a key role in testing the model performance. Shim et al. (2018) assessed how well the ionospheric models predict storm time f_oF_2 and TEC by considering quantities, such as TEC and f_oF_2 changes and percentage changes compared to quiet time background, at 12 selected midlatitude locations in the American and European-African longitude sectors. They found that the performance of the model varies with locations, even within a localized region like Europe, as well as with the metrics considered.

In this paper, we expand our previous assessment of modeled foF2 and TEC during 2013 March storm event (17 March, 2013) [*Shim et al*., 2018] to track improvement of the models and to investigate impacts of forcings from the lower atmosphere below and from the magnetosphere above on the performance of IT models. For this study, we evaluate the updated version of the coupled IT models available at the CCMC [*Webb et al*., 2009] since our previous study [*Shim et al*., 2018]: CTIPe version 4.1 and GITM version 21.11. However, the other types of models such as empirical models, stand-alone ionospheric models, and data assimilation models are not included. In addition, for the first time, simulations of NCAR WACCM-X 2.2 are included in our assessment. We also included a simulation of NCAR TIE-GCM 2 to compare with results from WACCM-X 2.2. For TEC prediction, we compare a weighted mean of the ensemble of all 13 simulations (ensemble average), including 8 simulations from our previous study with individual simulations to assess ensemble forecast capability. In Section 2, we briefly describe observations, models, and metrics used for this study. Section 3 presents the results of model-data comparisons and performance of the models are presented. Section 4 shows comparisons of ensemble of TEC predictions with the individual simulations based on the skill scores used in this study. Finally, we summarize and conclude in Section 5.

2. Methodology

2.1 Observations and Metrics

We use the foF2 and TEC measurements at 12 ionosonde stations selected in middle latitudes: 8 northern hemisphere (NH) stations in the US (Millstone Hill, Idaho national Lab, Boulder, and Eglin AFB) and Europe (Chilton, Pruhonice, Ebre, and Athens) and 4 southern hemisphere (SH) stations in South America (Port Stanley) and South Africa (Louisvale, Hermanus, and Grahamstown) (Figure 1 and Table 1 in *Shim et al.* [2018] for details). The foF2 and GNSS vertical TEC (vTEC) data are provided by Global Ionosphere Radio Observatory (GIRO) (http://giro.uml.edu/) [*Reinisch and Galkin*, 2011] and by MIT Haystack Observatory (http://cedar.openmadrigal.org/, http://cedar.openmadrigal.org/cgi-bin/gSimpleUIAccessData.py) [*Rideout and Coster*, 2006], respectively.

Table 1 shows the quantities and skill scores calculated for model-data comparison. To remove potential systematic uncertainties in the models and observations and baseline differences among the models and between models and observations, we use the shifted values and changes from their own quiet-time background values (e.g., shifted TEC (TEC*) = TEC (UT) on a particular DOY – median (UT) of TEC for 30 days centered on the storm date). Furthermore, using these quantities likely reduce the impacts of differing upper boundaries for TEC calculations, since the plasmaspheric TEC variations with geomagnetic activity are negligible in middle latitudes [*Shim et al*., 2017b].

To measure how well the observed and modeled values are linearly correlated (in phase) with each other and how different the values are on average over the time interval considered, CC and RMSE are calculated, respectively, for the error values below 95th percentile. We also calculate Yield and timing error to measure the models' capability to capture peak disturbances during the storm. For more detailed information on the quantities and skill scores used for the study, refer to Section 2 in *Shim et al.* [2018].

2.2 Models and Simulations

The simulations used in this study are obtained from the updated and newly incorporated coupled ionosphere-thermosphere models available at the CCMC [*Webb et al.*, 2009] since our previous study [*Shim et al.*, 2018]: CTIPe 4.1, GITM 21.11 and WACCM-X 2.2. The WACCM-X 2.2 simulations are provided by NCAR HAO. The WACCM-X version 2 [*Liu et al.*, 2018] is a comprehensive numerical model that extends the atmospheric component model of the NCAR Community Earth System Model (CESM) [*Hurrell et al*., 2013] into the thermosphere up to 500–700 km altitude. WACCM-X is uniquely capable of being run in a configuration where the atmosphere is coupled to active or prescribed ocean, sea ice, and land components, enabling studies of thermospheric and ionospheric weather and climate. WACCM-X version 2 is based upon WACCM version 6 [*Gettelman et al.*, 2019] with a top boundary of ~130 km, which is built upon the Community Atmosphere Model (CAM) version 6 having a top boundary of ~40 km. WACCM-X 2.2 includes WACCM6 physics for middle atmosphere and lower thermosphere as well as CAM6 physics for the troposphere and the lower stratosphere, and it fully incorporates the electrodynamical processes related to low-to mid-latitude wind dynamo that is implemented in the NCAR TIE-GCM. For this study, two specified-dynamics (SD) WACCM-X 2.2 simulations with different high-latitude electrostatic potential models [*Heelis et al.*, 1982; *Weimer*, 2005] are used. The SD simulations are carried out by constraining the model's lower atmospheric neutral dynamics using meteorological reanalysis data. The constraining process is achieved by nudging the model towards MERRA-2 (Modern Era Retrospective Analysis for Research and Applications, Version 2) data [*Gelaro et al.,* 2017] below around the altitude of 50 km in a way presented by *Brakebusch et al*. [2013].

The resulting WACCM-X simulations are compared with the simulations of TIE-GCM. The comparisons between WACCM-X and TIE-GCM simulations will show differences and similarities in modeling capabilities between whole atmosphere modeling and ionosphere-thermosphere modeling with a specified low-boundary forcing (e.g., Global Scale Wave Model (GSWM) [*Hagan et al.,* 1999] used for this study).

Table 2 shows the version of the models, input data used for the simulations, and models used for lower boundary forcing and high latitude electrodynamics. We utilized unique model setting identifiers to distinguish the current simulations from those used in our previous studies [*Shim et al.*, 2011, 2012, 2014, 2017a, 2018]. Additional information for the models and model setting identifiers is available in *Shim et al*. [2011] (Refer to all references therein) and at https://ccmc.gsfc.nasa.gov/support/GEM_metrics_08/tags_list.php

To investigate improvement in foF2 and TEC predictions of the updated versions of CTIPE (12_CTIPE) and GITM (7_GITM), the simulations of the old versions of the models (11_CTIPE and 6_GITM) from our previous study are included. The comparison will be focused on the comparison between the simulations obtained from the same model. As for TIE-GCM, 12_TIE-GCM (run at 2.5° resolution) is presented for this study, but the comparison between 11_TIE_GCM and 12_TIE-GCM was not included in this study because the only difference between the two is horizontal resolution (5°lat.×5°long. vs 2.5°lat.×2.5°long.).

We should take note of the difference between the simulations obtained from the same model that influence foF2 and TEC responses to geomagnetic storms. For two CTIPe runs, different lower atmospheric tides were specified: 11_CTIPE was driven by the imposed migrating semidiurnal (2,2), (2,3), (2,4), (2,5), and diurnal (1,1) tidal modes, while 12_CTIPE was run with monthly mean spectrum of tides obtained from WAM (Whole Atmosphere Model) [*Akmaev et al.,* 2011, *Fuller-Rowell et al.,* 2010]. For two GITM simulations, 7_GITM used Fang's auroral precipitation [*Fang et al.,* 2013], while 6_GITM used Ovation model [*Newell et al.,* 2009; 2011]. For two WACCM-X simulations, Heelis and Weimer2005 electric potential models were used for 3_WACCM-X and 4_WACCM-X, respectively. 12_TIEGCM was driven by Weimer2005 electric potential model and GSWM.

3. Performance of the Models in Predictions of foF2 and vTEC on 17 March 2013

Most simulations newly added for this study show similar behavior to those used in *Shim et al.* [2018], in predicting foF2 and TEC during the storm. For example, the simulations are not able to reproduce (1) the difference between eastern and western parts of the North American sector (e.g., TEC increases at Millstone Hill but decreases at Idaho and Boulder around 20UT), and (2) different responses between foF2 (negligible changes) and TEC (noticeable increase) found in European (Chilton) and South-African (Grahamstown) stations (See Figure 4 of Shim et al. [2018] for reference). However, compared to other simulations, 4_WACCM-X driven by *Weimer* (2005) high latitude electric potential model captures relatively well the two differences in TEC and foF2 described above (Figure S1 in supporting information).

Figure 1 shows scatter plots of the observed (*x* axis) and modeled (*y* axis) shifted foF2 and TEC, and percentage change of foF2 and TEC during the storm (03/17/2013) for all 12 locations grouped into 4 sectors: North America (NA, green), Europe (EU, blue), South Africa (SAF, red), and South America (SAM, black). First of all, the qualitative comparison between the simulations from the same model can be summarized as follows. 11_CTIPE/12_CTIPE tends to underestimate foF2 for both quiet and disturbed conditions, but 12_CTIPE predicts much better both foF2 and TEC during the storm than 11_CTIPE. 6_GITM and 7_GITM underestimate foF2 and TEC for all cases and show relatively small response to the storm compared to the other simulations. 12_TIE-GCM and WACCM-Xs produce similar foF2 and TEC changes during the storm. All three simulations give *substantial underestimation of TEC in SAF.* 12_TIE-GCM and *3_WACCM-X* produce larger overestimation of foF2 and TEC in NA sector than *4_WACCM-X. 4_WACCM-X* shows substantial improvement in the TEC overestimation in NA. 3_WACCM-X, of which the high latitude electric potential is specified by Heelis et al. [1982], tends to overestimate foF2 and TEC compared with 4_WACCM-X. 3_WACCM-X and 4_WACCM-X produce better quiet time foF2 and TEC than 12_TIEGCM does and capture wave-like small increases in foF2 and TEC at Idaho National Lab around 10–11UT (2–3 LT) (Figure S1 in supporting information).

As shown for 6_GITM and 11_CTIPE in *Shim et al.* [2018], the modeled foF2 values of 7_GITM and 12_CTIPE better agrees with the observed ones when they are shifted by subtracting the minimum of 30-day median (see Figure S2 in supporting information, *Shim et al.* [2018]). Most foF2 and TEC data points of 7_GITM and 12_CTIPE before shifting are below and above the line with slope 1 (black solid line), respectively. This indicates that 7_GITM underestimates foF2 and TEC like 6_GITM, while 12_CTIPE overestimates them. The models that tend to underestimate foF2, such as 6_GITM, 7_GITM and 11_CTIPE, seem to unable to produce foF2* larger than about 7 MHz, and underestimate TEC* being less than about 20 TECU during the storm as reported in *Shim et al.* [2018]. 12_TIE-GCM and WACCM-Xs show similar distribution of the data points after shifting foF2 and TEC with a tendency to underestimate foF2 and TEC in the South Africa region.

The modeled dfoF2[%] and dTEC[%] show less agreement with the observed values than the modeled foF2* and TEC* do. The data points in the 2nd quadrant (top left) and the 4th quadrant (bottom right) indicate that the modeled and observed percentage changes are in opposite sign. 7_GITM and 3_WACCM-X have more data points in the 2nd quadrant for dfoF2[%] prediction than 6_GITM and 4_WACCM-X, respectively. Like most simulations used in our previous evaluation [*Shim et al.* 2018], 12_CTIPE and 7_GITM do not appear to reproduce the large dTEC[%] (about 200 %) at Port Stanley in SAM. However, 12_TIE-GCM and WACCM-Xs better produce the enhancement in TEC percentage change. Compared to 4_WACCM-X and 12_TIE-GCM, 3_WACCM-X overestimates dTEC[%] especially in NA and EU regions. 12_CTIPE and 6_GITM have more data points of overestimated dTEC[%] in SAF than 11_CTIPE and 7_GITM, respectively.

From now on, foF2 and TEC will represent shifted foF2 (foF2*) and shifted TEC (TEC*), respectively.

3.1 Correlation Coefficient (CC)

We first calculate correlation coefficient (CC) between the modeled and observed foF2 and TEC for DOY 076 (17 March, 2013) for quantitative assessment of the model performance of TEC and foF2 predictions. In Figure 2, the CCs for each simulation are presented for foF2 in the left panel and for TEC in the right panel. For each simulation, four CC values are displayed. First three of the values correspond to the average CC over Europe (EU), North America (NA), Southern Hemisphere (SH refers to SAF and SAM combined), and the last one is the average of all 12 locations. The modeled foF2 and TEC (blue dots) are highly correlated with the observed values. The average CC values over all 12 locations for both foF2 and TEC are about 0.8–0.95, but the average CCs for their changes are much smaller. For example, the CCs for TEC changes (dTEC) are 0.5–0.6 and even smaller for foF2. The modeled foF2 changes (green), percentage changes (red) and normalized percentage changes (black only applicable for TEC) are much less correlated (closer to uncorrelated) with the observed values (about 0.1 < average CC < 0.4). There is no big difference between dTEC[%] and dTEC[%]_norm based on the average values for each simulation as reported in *Shim et al.* [2018].

Note that the CC values for the changes and percentage changes of foF2 and TEC are highly dependent on locations. Most simulations, except for 12_CTIPE and GITMs, show lower CC for dfoF2 and dTEC in NA. It seems to be caused by the decreases of foF2 and TEC during the storm (negative phase) in the western parts of NA that are not captured well. GITMs show the negative phase well although it underestimated the magnitude of the change. The CCs for the percentage changes of foF2 and TEC are particularly small for CTIPEs and GITMs.

11_CTIPE's foF2 and TEC averaged over 12 locations are slightly better correlated with the observed values than 12_CTIPE. However, the changes and percentage changes of foF2 and TEC from 12_CTIPE are better correlated with the observed values than 11_CTIPE's values in most regions. Although the two GITMs produce similar CCs, 7_GITM shows better CC in NA regions for dfoF2, dfoF2[%], dTEC[%], and n_dTEC[%], while 6_GITM shows better CC for foF2 and dTEC. WACCM-Xs perform better than 12_TIE_GCM for all the considered quantities based on the average except for dTEC. WACCM-Xs perform similar to each other.

Close inspection of Figures. 1 and 2 indicates that a linearity between CTIPE and observations is improved in the newer version of CTIPE (12_CTIPE), but 12_CTIPE gives more scattered distribution around a linear relation (Fig. 1), which seems to lead to the lower CC in 12_CTIPE than in 11_CTIPE. 7_GITM exhibits a slight improvement in a linearity between the model and observations (Fig. 1), but this improvement is not clearly seen in the correlation analysis (Fig. 2). For 12_TIEGCM and WACCM-Xs, both a linearity between the models and observations (Fig. 1) and CCs (Fig. 2) demonstrate that the model performances are overall improved in WACCM-Xs compared with TIEGCM. In terms of the model-observation linearity, 4_WACCMX is somewhat better than 3_WACCMX (Fig. 1), but their CCs seems comparable to each other (Fig. 2).

3.2 Root Mean Square Error (RMSE)

Figure 3 shows RMSE of foF2 and dfoF2 in the left panel, and TEC and dTEC in the right panel. For foF2 (blue) and dfoF2 (green) predictions, based on the average RMSE values, the

RMSEs from the updated version (12_CTIPE and 7_GITM) are about 1.5 MHz for foF2 and about 1 MHz for dfof2, and they are slightly lower than RMSEs in their old versions. 12_CTIPE shows improvement in foF2 in SH and dfoF2 in NA and EU compared to 11_CTIPE. 7_GITM performs better in foF2 and dfoF2 in EU and SH than 6_GITM. 4_WACCM-X has smaller RMSE (~1 MHz) than 3_WACCM-X and 12_TIE-GCM (~1.3 MHz for dfoF2 and ~2 MHz for foF2).

12_CTIPE is better in TEC prediction than 11_CTIPE, while the opposite holds true for dTEC prediction. The two GITMs' average RMSE values for TEC and dTEC predictions are similar to each other, about 9 TECU for TEC and 5 TECU for dTEC. Like foF2 and dfoF2 prediction, 4_WACCM-X has smaller RMSE (~ 5 TECU for TEC and 4 TECU for dTEC) than 12_TIE-GCM and 3_WACCM-X (~6 TECU).

As seen in *Shim et al.* [2018], RMSE is highly variable with location. Most simulations appear to predict foF2 and/or TEC better in NA and worse in SH (except for 12_TIE-GCM for foF2 and 12_CTIPE for TEC). Both 11_CTIPE and GITMs tend to perform better in NA for dTEC, while WACCM-Xs show the opposite tendency for dfoF2 and dTEC. 7_GITM and 4_WACCM-X shows the least RMSE dependence on location for dfoF2 and for dTEC, respectively, among seven simulations.

Figure 4 shows the RMSE of percentage changes of foF2 (blue) and TEC (red) and normalized percentage changes of TEC (black). The two CTIPEs produce the similar RMSE for dTEC[%], but 12_CTIPE and 11_CTIPE produce lower RMSE for dfoF2[%] and dTEC[%]_norm, respectively. For all three percentage changes of dfoF2[%], dTEC[%], and dTEC[%]_norm, 7_GITM seems to perform better than 6_GITM based on the average RMSEs over the 12 locations. 4_WACCM-X and 12_TIE-GCM perform very similarly for dfoF2[%] and dTEC[%] and better than 3_WACCM-X.

Difference in the performance among locations is more noticeable in dTEC[%] and dTEC[%]_norm than in dfoF2[%] as found in *Shim et al.* [2018]. All simulations, except 6_GITM, produce lower RMSE of dTEC[%] in NA and higher in SH region. This tendency remains the same for dTEC[%]_norm with the exception of 3_WACCM-X, which has lower RMSE for dTEC[%]_norm in SH. For 3_WACCM-X, the higher RMSE for dTEC[%] and the lower RMSE for dTEC[%]_norm in SH than in NA are probably due to the normalization factor, standard deviation of dTEC[%] in the locations.

3.3 Yield and Timing Error (TE)

To measure how well the models capture the degree of TEC and foF2 disturbances during the main phase, Yield and Timing Error (TE) of dfoF2[%], dTEC[%], and dTEC[%]_norm are calculated. *Shim et al.* [2018] considered two time intervals, 06–15UT and 15–22UT, when peaks are observed in most of 12 locations. In each time interval, we calculate one Yield value and one TE value. Definitions of Yield and TE are presented in Table 1.

In each sector, average Yield and TE are calculated over the number of stations where the model correctly predicts the storm phase, i.e., Yield is positive. Table 3 shows the total number of stations where the models show correct storm phase, either positive or negative. The numbers in bold are the higher values between the simulations compared. 12_CTIPE predicts the storm phase better for dTEC[%] than 11_CTIPE, but 11_CTIPE predicts better for dfoF2[%] than 12_CTIPE. 7_GITM is improved in predicting the storm phase of dfoF2[%], while 6_GITM predicts better the storm phase of dTEC[%]. 4_WACCM-X, compared to 12_TIE-GCM and

3_WACCM-X, is better for predicting the phase of dfoF2[%] and worse for predicting that of dTEC[%].

Figure 5 shows average Yield (left) and average of absolute values of TE (right) over the two time intervals: dfoF2[%] in blue, dTEC[%] in red, and dTEC[%]_norm in black. Concerning the average of all 12 locations, 12_CTIPE appears to overestimate peak values of dTEC[%] and dTEC[%]_norm with larger variation with location (e.g., ~1 < Yield of dTEC[%]_norm < ~2.5) than 11_CTIPE, of which Yield is less than 1 for all three quantities of percentage changes (e.g., 0.7 < Yield of dTEC[%]_norm < 0.9). Yields of 12_CTIPE for dTEC[%] and dTEC[%]_norm are closer to 1 in NA. GITMs produce similar ratios based on the average over all locations, but 7_GITM shows smaller differences in Yield among locations (e.g., ~0.5 < Yield of dTEC[%]_norm < ~1) than 6_GITM (e.g., 0.5 < Yield of dTEC[%]_norm < ~2.5). In terms of average Yield, 12 TIE-GCM and two WACCM-Xs tend to overestimate the peak values and show similar performance, although 12_TIE-GCM's ratios are closer to 1 than those of WACCM-Xs. 3_WACCM-X shows larger variation in Yield among locations (e.g., ~0.9 < Yield of dTEC[%]_norm < ~2.7) than 12_TIE-GCM and 4_WACCM-X (e.g., ~1.7 < Yield of dTEC[%]_norm < ~2.3).

Average Timing Errors of dfoF2[%] and dTEC[%]_norm are between 1 and 2 hours, and TE of dTEC[%] are about 0.8–1.5 hours. With respect to the average TE, 12_CTIPE has smaller TE (~1 hr) than 11_CTIPE (about 1.5 hr) for all three percentage changes with less location dependence as well. 7_GITM's three TEs are about 1.5 hrs, while 6_GITM's TEs of dfoF2[%], dTEC[%] and dTEC[%]_norm are ~1, ~1.4, and ~2 hrs, respectively. 12 TIE-GCM has smaller TE for dfoF2[%] and 3_WACCM-X has smaller TE for dTEC[%] and dTEC[%]_norm, however 3_WACCM-X show larger location dependence of TE for dTEC[%]_norm and dfoF2[%].

4. Ensemble of TEC obtained from13 simulations

The linearity check, RMSE, and CC between model results and observations for shifted foF2 and TEC and their relative changes indicate that the newer versions of the models (i.e., 12_CTIPE, 7_GITM and 4_WACCM-X) produces the better results. From the viewpoints of correct prediction of storm phases (Table 3), Yields, and TEs (Fig. 5), however, there is no one best simulation for all locations, and the performance of model varies with locations as well as the Yields and TE.

The differences in performance among the simulations could be caused by inherent differences among the models or by a combination of different input data and different models used for lower boundary forcing and high-latitude electrodynamics. Even different data assimilation models for the same weather condition can yield different results, due to numerous reasons (e.g., the use of different background weather models, spatial/temporal resolutions, assimilation methods, and data error analyses), even if the same data are assimilated [*Schunk et al.*, 2021]. The common way to handle these differences is to use model ensembles and the use of ensembles enables estimations of the certainty of results. Thus, we used a weighted mean of the ensemble of all 13 simulations including 8 simulations from our previous study (*Shim et al.,* 2018) for TEC, dTEC and dTEC[%] to compare the ensemble average with the individual simulations. To get the weighted mean ($\bar{x} = \sum w_i x_i / \sum w_i$), we used the RMSE of shifted TEC ($w_i = 1/\mathrm{RMSE}$).

Figure 6 is the same as Figure 1 but for the ensemble of the simulations (ENSEMBLE will be used as model setting ID) and a simulation (1_USU-GAIM) from a data assimilation model (DA), USU-GAIM. For TEC less than about 20 TECU, ENSEMBLE shows better agreement with GPS TEC than the individual simulations, including 1_USU-GAIM. However, as we can expect, ENSEMBLE underestimates TEC larger than about 30 TECU due to the tendency to underestimate TEC of many simulations as pointed out in Section 3 and *Shim et al.,* [2018]. For dTEC[%], ENSEMBLE appears to be correlated better with GPS dTEC[%] than the other simulations, although there are some underestimations in SAF, as well as in SAM with opposite prediction of the storm phase.

Figure 7 shows averaged CC and RMSE values over all 12 locations of 13 simulations, the ensemble of them, and the ensemble of 12 simulations excluding 1_USU-GAIM (ENSEMBLE_wo_DA). The simulations in Figure 7 (a) were arranged by the average of the three averaged CC values for TEC, dTEC and dTEC[%] from the smallest to the largest (closer to 1). In Figure 7 (b), the simulations were arranged by the average of the two averaged RMSEs for TEC and dTEC from the largest to the smallest. Based on the averaged CC and RMSE, ENSEMBLEs (ENSEMBLE and ENSEMBLE_wo_DA) of the simulations perform very similarly and outperform all 12 simulations but a data assimilation model, 1_USU-GAIM. However, ENSEMBLEs and 1_USU-GAIM do not show big difference in their performance. The differences in RMSE of TEC and dTEC between ENSEMBLE and 1_USU-GAIM are less than 0.5 and 0.1 TECU, respectively. For dTEC[%], ENSEMBLE performs slightly better than 1_USU-GAIM with about 1.5% lower RMSE. The fact that ENSEMBLEs are comparable to the data assimilation model 1_USU-GAIM indicates that the multi-model ensemble can be useful in forecasting the IT system, although this result is obtained from a single geomagnetic storm event.

Figure 8 shows Yield and Timing Error of dTEC[%] for all 13 simulations along with ENSEMBLE. The values correspond to the average over all 12 locations. Unlike CC and RMSE, ENSEMBLE does not outperform all physic-based coupled models in terms of Yield and TE, although the difference is small. ENSEMBLE underestimates Yield, while most of the simulations overestimate it, except 4_IRI and 11_CTIPE. 7 simulations from PB coupled IT models and 1_USU-GAIM produce Yield closer to 1 than ENSEMBLE does.

Timing Error of dTEC[%] of ENSEMBLE is about 1 hr, which is slightly larger than TE from 4 simulations from CTIPE and WACCM-X, but the difference from the smallest TE is less than 0.5 hr.

Regarding the averaged skill scores for all 12 locations, newly added five simulations in this study produce comparable TEC and TEC changes to the simulations from PB IT models used in our previous study. The simulations of newer versions of the models (12_CTIPE, 7_GITM and 4_WACCM-X) are found to give overall improved forecast results. Based on the averaged RMSE, the ensemble of simulations of the models' newer versions is comparable to 1_USU-GAIM and performs better than the ensemble of the simulations of old versions of models (11_CTIPE, 6_GITM and 12_TIE-GCM) (Table 4).

5. Summary and Conclusions

We expanded on our previous systematic assessment of modeled foF2 and TEC during 2013 March storm event (17 March, 2013) to track the improvement of the models and investigate impacts of forcings from the lower atmosphere and the magnetosphere, on the performance of ionosphere-thermosphere coupled models.

We evaluated simulations from upgraded models (CTIPe4.1 and GITM21.11) since our previous assessment and a whole atmosphere model (WACCM-X2.2). To compare with results from WACCM-X2.2, we also included a simulation of TIE-GCM2.0, of which the electrodynamic processes are implemented in WACCM-X 2.2. Furthermore, to evaluate TEC prediction of the simulations, we used a weighted mean of the ensemble of all 13 simulations including 8 simulations from our previous study to compare the ensemble average with the individual simulations.

For evaluation of the simulations, we used the exact same procedure with the same data set, same physical quantities, and same skill scores as our previous study [*Shim et al.*, 2018]. The skill scores were calculated for the three sectors, EU (Europe), NA (North America), and SH (Southern Hemisphere) to investigate the longitudinal and hemispheric dependence of the performance of the models.

From the five simulations used in the study, we also found the general behaviors of most simulations identified in *Shim et al.* [2018]: 1) tendency to underestimate storm-time enhancements of foF2 and TEC and not to reproduce large enhancements of dTEC[%] (e.g., about 200 % TEC increase at Port Stanley in the SAA region), 2) being unable to capture opposite responses to the storm in the eastern and western parts of NA, especially negative phase (except for GITM), which is what in part causes lower CC in NA, 3) tendency to predict foF2 and/or TEC better in NA and worse in SH with respect to RMSE. However, it was found that 12_TIE-GCM and WACCM-Xs better produce the large TEC percentage changes at Port Stanley in SAM. Based on the averaged skill scores for all 12 locations, the five simulations used in this study show skill scores better or comparable to those of the simulations from PB IT models used in our previous study.

Compared to 11_CTIPE (obtained from CTIPe3.2), 12_CTIPE (from CTIPe4.1) driven by tides from WAM tends to overestimate foF2 and TEC for both quiet and disturbed conditions and predicts better TEC peaks during the storm. For more cases, 12_CTIPE performs largely better than 11_CTIPE based on the average scores. 12_CTIPE predicts the storm phase better for dTEC[%], but 11_CTIPE does better for dfoF2[%]. 12_CTIPE appears to overestimate peak values of dTEC[%] and dTEC[%]_norm, while 11_CTIPE produces Yield less than 1.

The two GITMs, 7_GITM with Fang's auroral precipitation and 6_GITM with Ovation model, underestimate foF2 and TEC for all cases and show relatively small response to the storm compared to the other simulations that do not appear to reproduce the large dTEC[%] (about 200 % increase at Port Stanley in SAM). 7_GITM and 6_GITM perform very similarly for most cases with similar skill scores. However, 7_GITM shows better CC for most quantities except for dTEC, and lower RMSEs and Yield closer to 1 for most regions and quantities considered. 7_GITM shows the least RMSE dependence on location for dfoF2 among the other simulations.

Comparing two WACCM-Xs and 12_TIE-GCM, the two WACCM-Xs, 3_WACCM-X with Heelis high latitude electric potential model and 4_WACCM-X with Weimer 2005, predict quiet time foF2 and TEC better than 12_TIE-GCM. During the storm, 12_TIE-GCM and 4_WACCM-X produce similar foF2 and TEC in NA sector, while 3_WACCM-X tends to overestimate them and produces larger changes in foF2 and TEC. In most cases, WACCM-Xs and 12_TIE_GCM perform similarly in terms of average values of skill scores, but 3_WACCM-X and/or 4_WACCM-X perform better than 12_TIE-GCM except for Yield of percentage changes. 4_WACCM-X slightly outperforms 3_WACCMX for all cases but not for TE for percentage changes.

Our findings suggest that the newer versions of the models (12_CTIPE, 7_GITM and 4_WACCM-X) with Weimer2005 electric potential model give overall improved forecast, and the performance of the models depends on forcing from the magnetosphere and also forcing from the lower atmosphere even during storms.

For TEC, dTEC and dTEC[%], our results indicate that the ensemble of all 13 simulations (ENSEMBLE), including 8 simulations from our previous study (*Shim et al.,* 2018) is comparable to the data assimilation model (1_USU-GAIM) with differences in skill score less than 3% and 6% for CC and RMSE, respectively. However, ENSEMBLE underestimates Yield (0.73) while 7 simulations from PB coupled IT models and 1_USU-GAIM produce Yield closer to 1. Timing Error of dTEC[%] of ENSEMBLE is about 1 hr, but the difference from the smallest TE of the simulations is less than 0.5 hr. In addition, based on RMSE, the ensemble of the newer versions of the models (12_CTIPE, 7_GITM and 4_WACCM-X) is comparable to 1_USU-GAIM.

To advance our understanding of the ionosphere-thermosphere system requires significant efforts to improve the capability of numerical models along with the scope of observations [*Heelis and Maute*, 2020]. There have been recent new developments of theoretical models, including AMGeO (Assimilative Mapping of Geospace Observations) for High-Latitude Ionospheric Electrodynamics [*Matsuo*, 2020] and MAGE geospace model that couples the Grid Agnostic MHD for Extended Research Applications (GAMERA) global MHD model of the magnetosphere (Sorathia et al., 2020; Zhang et al., 2019), the Rice Convec-tion Model (RCM) model of the ring current (Toffoletto et al., 2003), TIEGCM of the upper atmosphere and the RE-developed Magnetosphere-Ionosphere Coupler/Solver (REMIX) (Merkin & Lyon, 2010). These models will be available soon to the public through CCMC, and then the modeling capability will help us better understand the processes responsible for the observed characteristics and features during disturbed conditions. In addition, CCMC will also provide users with the capability to run PB IT models with various combination of models for lower atmospheric forcing and for magnetosphere forcing, which enable us to research further the impacts of the forcings on the IT system.

The findings of this study will provide a baseline for future validation studies using new models and improved models, along with earlier results [*Shim et al.*, 2011, 2012, 2014, 2017a, 2018] obtained through CEDAR ETI, GEM-CEDAR Modeling Challenges, and the international effort, "International Forum for Space Weather Modeling Capabilities Assessment". We will extend our study to include more geomagnetic storm events to investigate differences and similarities in the performance of the models. In addition, we will also include foF2 and TEC predictions for the high- and low-latitude regions.

Acknowledgement

This work was supported by Korea Polar Research Institute (KOPRI) grant funded by the Ministry of Oceans and Fisheries (KOPRI PE22020) and basic research funding from the Korea Astronomy and Space Science Institute (KASI) (KASI2022185009). The vertical TEC data were provided by MIT Haystack Observatory and can be obtained through CEDAR Madrigal database (http://cedar.openmadrigal.org). We thank the operators of the digisondes for sharing their data through http://giro.uml.edu/. Data from the South African Ionosonde network is made available through the South African National Space Agency (SANSA), who are acknowledged for facilitating and coordinating the continued availability of data. This work is supported by grants from the National Science Foundation (NSF) Space Weather Program. This model validation study is supported by the Community Coordinated Modeling Center (CCMC) at the Goddard Space Flight Center. Data processing and research at MIT Haystack Observatory are supported by cooperative agreement AGS-1242204 between the U.S. National Science Foundation and the

Massachusetts Institute of Technology. The National Center for Atmospheric Research is sponsored by the National Science Foundation. Model output and observational data used for the study will be permanently posted at the CCMC website (http://ccmc.gsfc.nasa.gov) and provided as a resource for the space science community to use in the future.

References

Akmaev, R. A. (2011). Whole atmosphere modeling: Connecting terrestrial and space weather. *Reviews of Geophys. 49*, RG4004. 390 https://doi.org/10.1029/2011RG000364

Brakebusch, M., Randall, C. E., Kinnison, D. E., Tilmes, S., Santee, M. L., and Manney, G. L. (2013) Evaluation of Whole Atmosphere Community Climate Model simulations of ozone during Arctic winter 2004–2005, J. Geophys. Res., 118, 2673–2688, https://doi.org/10.1002/jgrd.50226

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

Chamberlin, P. C., Woods, T. N., & Eparvier, F. G. (2007). Flare Irradiance Spectral Model (FISM): Daily component algorithms and results. *Space Weather, 5,* S07005. https://doi.org/10.1029/2007SW000316

Codrescu, M. V., T. J. Fuller-Rowell, J. C. Foster, J. M. Holt, and S. J. Cariglia, (2000), Electric field variability associated with the Millstone Hill electric field model, *J. Geophys. Res*., 105, 5265–5273, doi:10.1029/1999JA900463.

Fang, X., D. Lummerzheim, and C. H. Jackman (2013), Proton impact ionization and a fast calculation method, J. Geophys. Res. Space Physics, 118, 5369–5378, doi:10.1002/jgra.50484.

Fuller -Rowell, T. J., and D. S. Evans, (1987), Height-Integrated Pedersen and Hall Conductivity Patterns Inferred From the TIROS-NOAA Satellite Data, *J. Geophys. Res.*, 92(A7), 7606–7618.

Fuller-Rowell, T., Wu, F., Akmaev, R., Fang, T.-W., & Araujo-Pradere, E. (2010). A whole atmosphere model simulation of the impact of a sudden stratospheric warming on thermosphere dynamics and electrodynamics. *Journal of Geophysical Research*, *115*, A00G08. https://doi.org/10.1029/2010JA015524

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *Journal of Climate*, 30(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., et al.(2019). The whole atmosphere community climate model version 6 (WACCM6), *Journal of Geophysical Research: Atmospheres*, *124*, 12,380–12,403. https://doi.org/10.1029/2019JD030943.

Hagan, M. E., M. D. Burrage, J. M. Forbes, J. Hackney, W. J. Randel, and X. Zhang, (1999), GSWM-98: results for migrating solar tides. *J. Geophys. Res.* 104: 6813–6828.

Hedin, A. E. (1991), Extension of the MSIS thermospheric model into the middle and lower atmosphere, *J. Geophys. Res*., 96, 1159–1172.

Heelis, R. A., J. K. Lowell, and R. W. Spiro, (1982), A Model of the High-Latitude Ionospheric Convection Pattern, *J. Geophys. Res*. 87, 6339.

Heelis, R. A., & Maute, A. (2020). Challenges to understanding the Earth's ionosphere and thermosphere. *JGR: Space Physics*, *125*, https:// doi.org/10.1029/2019JA027497

Jin, H., Miyoshi, Y., Fujiwara, H., Shinagawa, H., Terada, K., Terada, N., et al. (2011). Vertical connection from the tropospheric activities to the ionospheric longitudinal structure simulated by a new Earth's whole atmosphere-ionosphere coupled model. *Journal of Geophysical Research*, *116*, A01316. https://doi.org/10.1029/2010JA015925

Kalafatoglu Eyiguler, E. C., Shim, J. S.,Kuznetsova, M. M., Kaymaz, Z.,Bowman, B. R., Codrescu, M. V., 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.

Liu, H.-L., Bardeen, C. G., Foster, B. T., Lauritzen, P., Liu, J., Lu, G., . . . Wang, W. (2018). Development and validation of the Whole Atmosphere Community Climate Model with thermosphere and ionosphere extension (WACCM-X 2.0), *Journal of Advances in Modeling Earth Systems*, 10. https://doi.org/10.1002/2017MS001232

Matsuo, T. (2020). Recent Progress on Inverse and Data Assimilation Procedure for High-Latitude Ionospheric Electrodynamics. In: Dunlop, M., Lühr, H. (eds) Ionospheric Multi-

Spacecraft Analysis Tools. ISSI Scientific Report Series, vol 17. Springer, Cham. https://doi.org/10.1007/978-3-030-26732-2_10

Merkin, V., & Lyon, J. (2010). Effects of the low-latitude ionospheric boundary condition on the global magnetosphere. *Journal of Geophysical Research*, *115*(A10). A10202. https://doi.org/10.1029/2010JA015461

Millward, G. H., I. C. F. Müller-Wodrag, A. D. Aylward, T. J. Fuller-Rowell, A. D. Richmond, and R. J. Moffett, (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, 24,733–24,744, doi:10.1029/2000JA000342.

Newell, P. T., T. Sotirelis, and S. Wing (2009), Diffuse, monoenergetic, and broadband aurora: The global precipitation budget, *J. Geophys. Res*., 114, A09207, doi: 10.1029/2009JA014326.

Newell, P.T., and J.W. Gjerloev (2011), Substorm and magnetosphere characteristic scales inferred from the SuperMAG auroral electrojet indices, *J. Geophys. Res*., 116, A12232, doi:10.1029/2011JA016936.

Rastäetter, L., et al., (2016), GEM-CEDAR Challenge: Poynting Flux at DMSP and modeled Joule Heat, *Space Weather*, *14*, 113–135, doi:10.1002/2015SW001238.

Reinisch, B., and I. Galkin, (2011). Global Ionospheric Radio Observatory (GIRO). *Earth, Planets, and Space. 63*. 377-381. 10.5047/eps.2011.03.001.

Richmond, A. D., E. C. Ridley and R. G. Roble, (1992), A Thermosphere/Ionosphere General Circulation Model with coupled electrodynamics, *Geophys. Res. Lett.*, **19**, 601-604.

Rideout, W., and A. Coster, (2006), Automated GPS processing for global total electron content data, GPS Solution, doi:10.1007/s10291-006-0029-5.

Ridley, A. J., Y. Deng, and G. Toth, (2006), The global ionosphere-thermosphere model, *J. Atmos. Sol. Terr. Phys*., 68, 839-864.

Roble, R. G., E. C. Ridley, A. D. Richmond, and R. E. Dickinson, (1988), A coupled thermosphere/ionosphere general circulation model, *Geophys. Res. Lett*., 15, 1325–1328, doi:10.1029/GL015i012p01325.

Scherliess, L., Tsagouri, I., Yizengaw, E., Bruinsma, S., Shim, J. S., Coster, A., and Retterer, J. M. (2019). The International Community Coordinated Modeling Center space weather modeling capabilities assessment: Overview of ionosphere/thermosphere activities. *Space Weather*, *17*. https:// doi.org/10.1029/2018SW002036

Schunk, R. W., Scherliess, L., Eccles, V., Gardner, L. C., Sojka, J. J., Zhu, L., et al. (2021). Challenges in specifying and predicting space weather. *Space Weather*, *19*, e2019SW002404. https:// doi.org/10.1029/2019SW002404

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., et al., (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*, S10004, doi:10.1029/2012SW000851.

Shim, J. S., et al., (2014), Systematic Evaluation of Ionosphere/Thermosphere (IT) Models: CEDAR Electrodynamics Thermosphere Ionosphere (ETI) Challenge (2009-2010), in *Modeling the Ionosphere-Thermosphere System*, AGU Geophysical Monograph Series.

Shim, J. S., Rastätter, L., Kuznetsova, M., Bilitza, D., Codrescu, M., Coster, A. J., … Zhu, L. (2017a). CEDAR-GEM challenge for systematic assessment of Ionosphere/thermosphere models in predicting TEC during the 2006 December storm event. *Space Weather, 15*, 1238–1256. https://doi.org/10.1002/ 2017SW001649

Shim, J. S., G. Jee, and L. Scherliess (2017b), Climatology of plasmaspheric total electron content obtained from Jason 1 satellite, J. Geophys. Res. Space Physics, 122, 1611–1623, doi:10.1002/2016JA023444.

Shim, J. S., Tsagouri, I., Goncharenko, L., Rastaetter, L., Kuznetsova, M., Bilitza, D., et al. (2018). Validation of ionospheric specifications during geomagnetic storms: TEC and foF2 during the 2013 March storm event. Space Weather, 16, 1686–1701. https://doi.org/10.1029/ 2018SW002034

Solomon, S. C., A. G. Burns, B. A. Emery, M. G. Mlynczak, L. Qian, W. Wang, D. R. Weimer, and M. Wiltberger (2012). Modeling studies of the impact of high-speed streams and co-rotating interaction regions on the thermosphere-ionosphere. *J. Geophys. Res.*, *117*, A00L11, doi:10.1029/2011JA017417

Sorathia, K., Merkin, V., Panov, E., Zhang, B., Lyon, J., Garretson, J., et al. (2020). Ballooning-interchange instability in the near-Earth plasma sheet and auroral beads: Global magnetospheric modeling at the limit of the MHD approximation. *Geophysical Research Letters*, *47*(14), e2020GL088227. https://doi.org/10.1029/2020GL088227

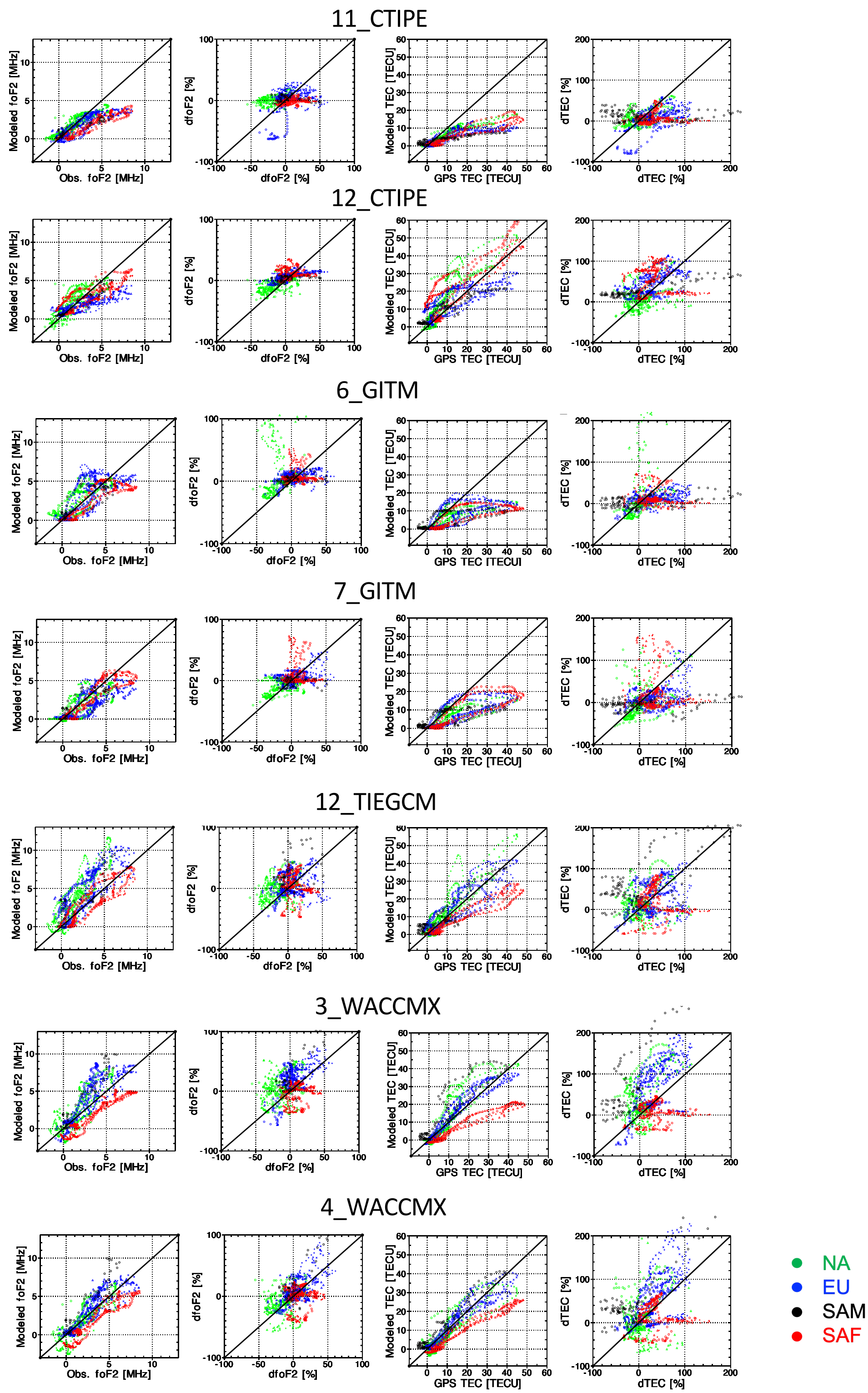
Tsagouri, I., Goncharenko, L., Shim, J. S., Belehaki, A., Buresova, D., & Kuznetsova, M. (2018). Assessment of current capabilities in modeling the ionospheric climatology for space weather applications: foF2 and hmF2. *Space Weather*, *16*, 1930–1945. https://doi.org/10.1029/2018SW002035

Toffoletto, F., Sazykin, S., Spiro, R., & Wolf, R. (2003). Inner magnetospheric modeling with the rice convection model. *Space Science Reviews*, *107*(1–2), 175–196. https://doi.org/10.1023/A:1025532008047

Webb, P. A., M. M. Kuznetsova, M. Hesse, L. Rastaetter, and A. Chulaki, (2009), Ionosphere-thermosphere models at the Community Coordinated Modeling Center, *Radio Sci*., 44, RS0A34, doi:10.1029/2008RS004108.

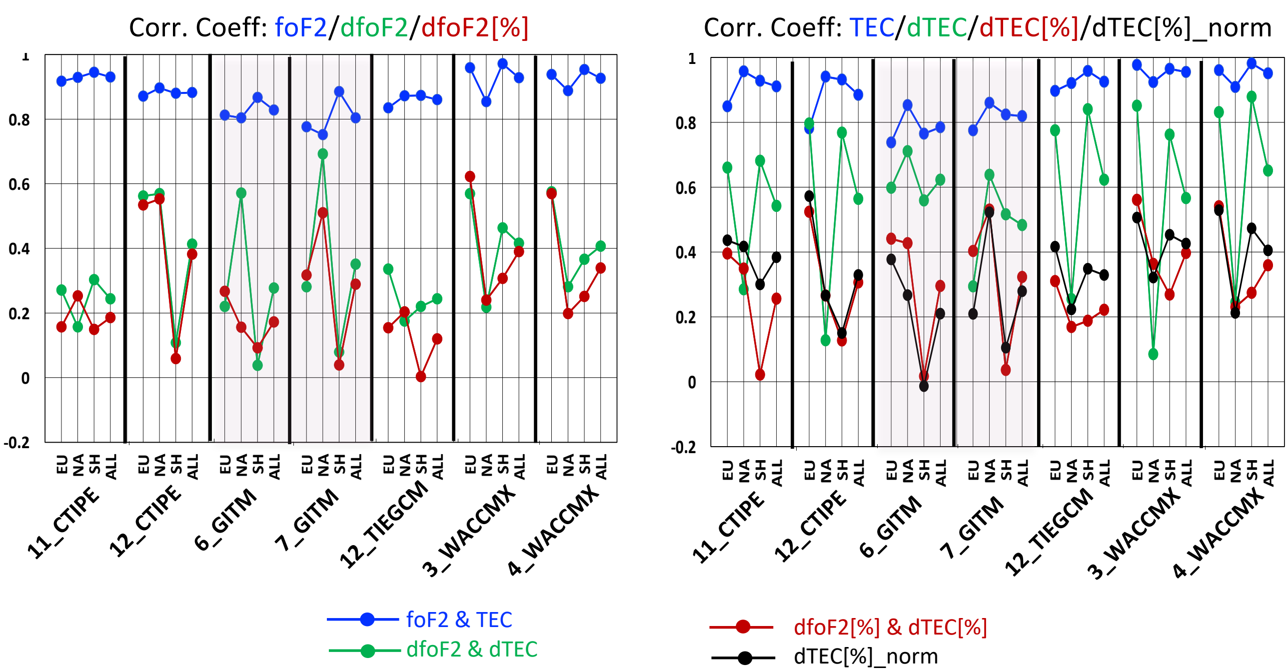
Weimer, D. R., (2005), Improved ionospheric electrodynamic models and application to calculating Joule heating rates, *J. Geophys. Res*., *110,* A05306, doi:10.1029/2004JA010884.

Zhang, B., Sorathia, K. A., Lyon, J. G., Merkin, V. G., Garretson, J. S., & Wiltberger, M. (2019). GAMERA: A three-dimensional finite-volume MHD solver for non-orthogonal curvilinear geometries. *The Astrophysical Journal Supplement Series*, *244*(1), 20. https://doi.org/10.3847/1538- 4365/ab3a4c



11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
Modeled foF2 [MHz]
Obs. foF2 [MHz]
dfoF2 [%]
Modeled TEC [TECU]
GPS TEC [TECU]
dTEC [%]
NA
EU
SAM
SAF

Figure 1



Corr. Coeff: foF2/dfoF2/dfoF2[%]
Corr. Coeff: TEC/dTEC/dTEC[%]/dTEC[%]_norm
1
0.8
0.6
0.4
0.2
0
-0.2
EU NA SH ALL
11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
foF2 & TEC
dfoF2 & dTEC
dfoF2[%] & dTEC[%]
dTEC[%]_norm

Figure 2

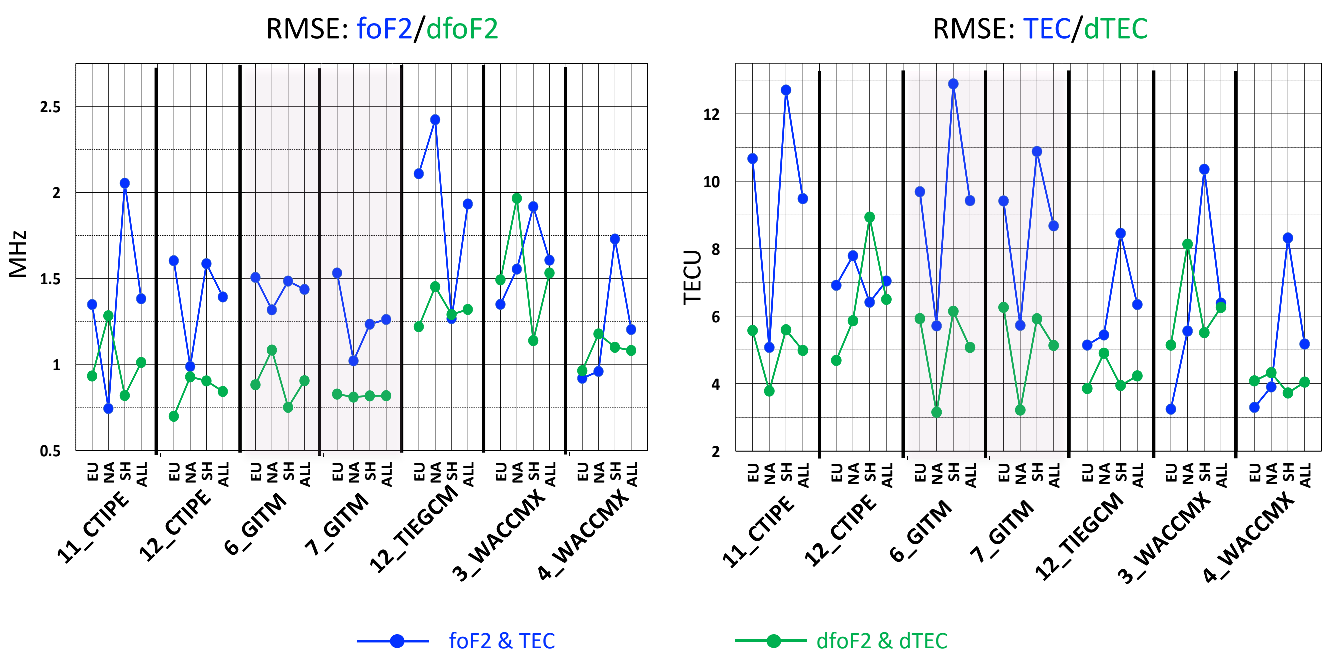

RMSE: foF2/dfoF2
RMSE: TEC/dTEC
MHz
TECU
2.5
2
1.5
1
0.5
12
10
8
6
4
2
EU
NA
SH
ALL
11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
foF2 & TEC
dfoF2 & dTEC

Figure 3

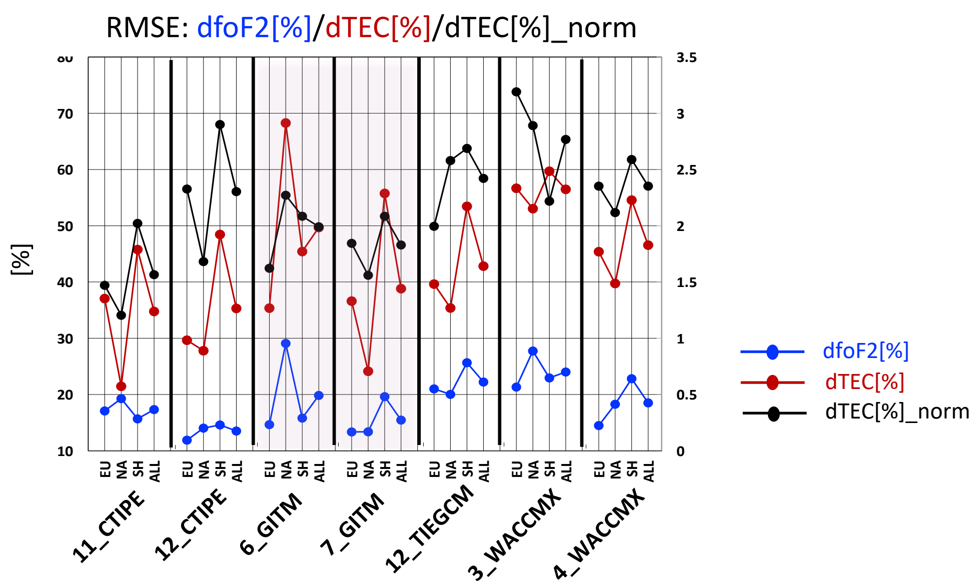
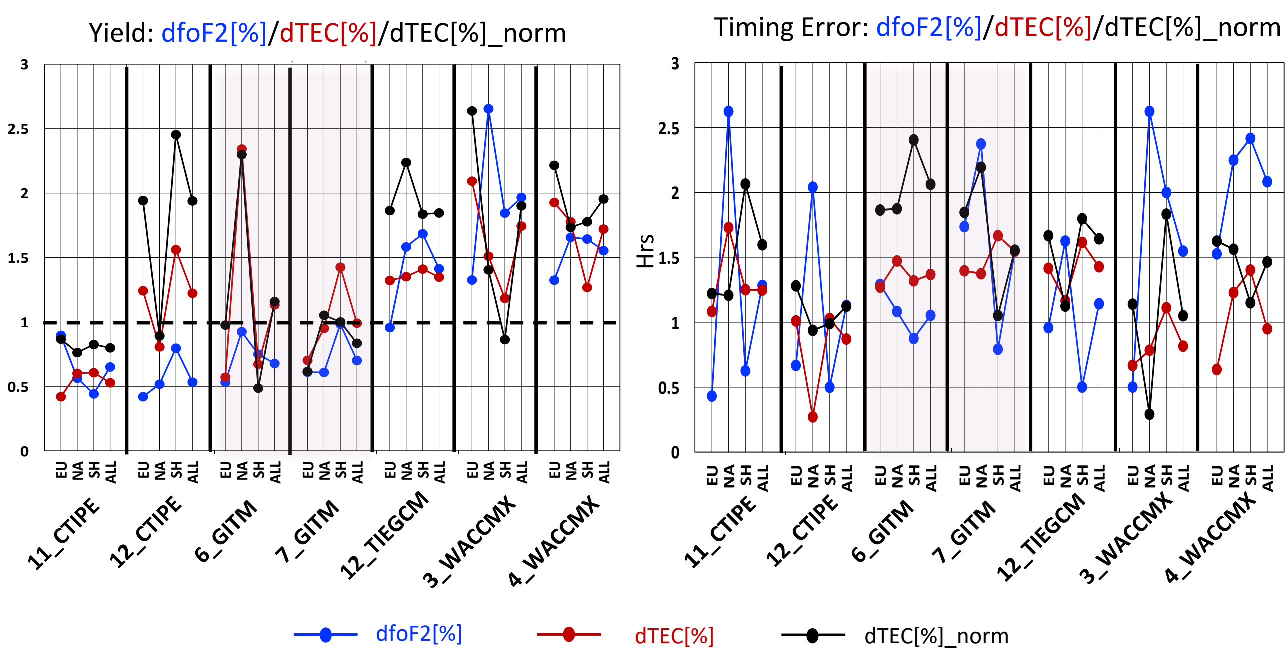

RMSE: dfoF2[%]/dTEC[%]/dTEC[%]_norm
[%]
80
70
60
50
40
30
20
10
3.5
3
2.5
2
1.5
1
0.5
0
EU
NA
SH
ALL
11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
dfoF2[%]
dTEC[%]
dTEC[%]_norm

Figure 4



Yield: dfoF2[%]/dTEC[%]/dTEC[%]_norm
Timing Error: dfoF2[%]/dTEC[%]/dTEC[%]_norm
Hrs
0
0.5
1
1.5
2
2.5
3
EU
NA
SH
ALL
11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
dfoF2[%]
dTEC[%]
dTEC[%]_norm

Figure 5

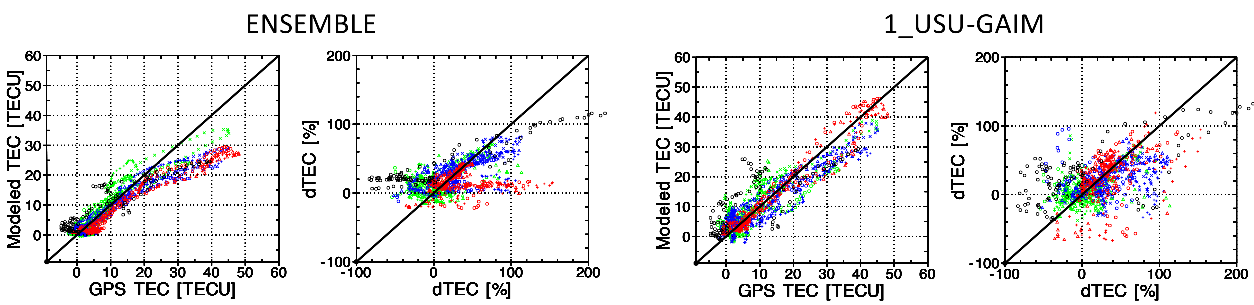

ENSEMBLE
1_USU-GAIM
Modeled TEC [TECU]
GPS TEC [TECU]
dTEC [%]
dTEC [%]
Modeled TEC [TECU]
GPS TEC [TECU]
dTEC [%]
dTEC [%]

Figure 6

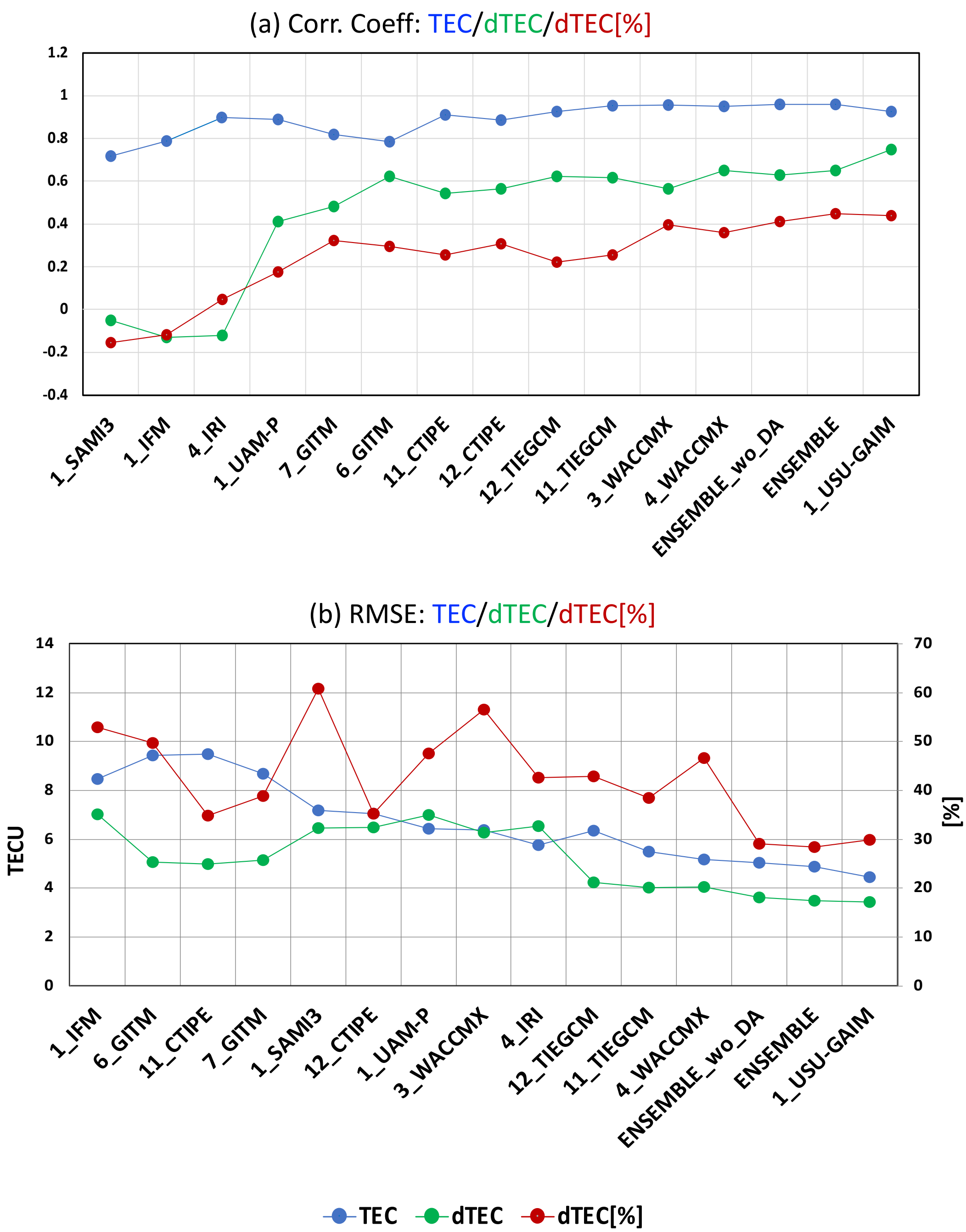
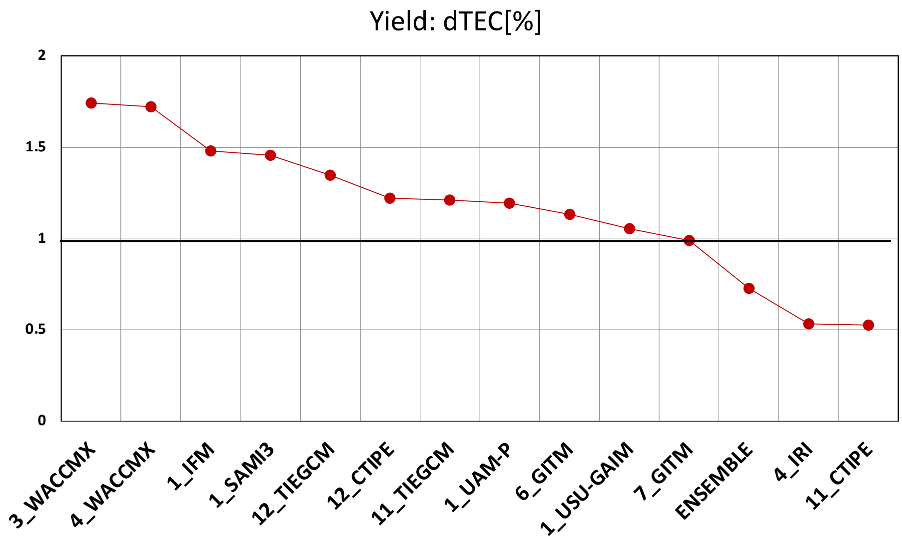

(a) Corr. Coeff: TEC/dTEC/dTEC[%]
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
1_SAMI3
1_IFM
4_IRI
1_UAM-P
7_GITM
6_GITM
11_CTIPE
12_CTIPE
12_TIEGCM
11_TIEGCM
3_WACCMX
4_WACCMX
ENSEMBLE_wo_DA
ENSEMBLE
1_USU-GAIM
(b) RMSE: TEC/dTEC/dTEC[%]
TECU
[%]
14
12
10
8
6
4
2
0
70
60
50
40
30
20
10
0
1_IFM
6_GITM
11_CTIPE
7_GITM
1_SAMI3
12_CTIPE
1_UAM-P
3_WACCMX
4_IRI
12_TIEGCM
11_TIEGCM
4_WACCMX
ENSEMBLE_wo_DA
ENSEMBLE
1_USU-GAIM
TEC
dTEC
dTEC[%]

Figure 7


Yield: dTEC[%]
2
1.5
1
0.5
0
3_WACCMX
4_WACCMX
1_IFM
1_SAMI3
12_TIEGCM
12_CTIPE
11_TIEGCM
1_UAM-P
6_GITM
1_USU-GAIM
7_GITM
ENSEMBLE
4_IRI
11_CTIPE

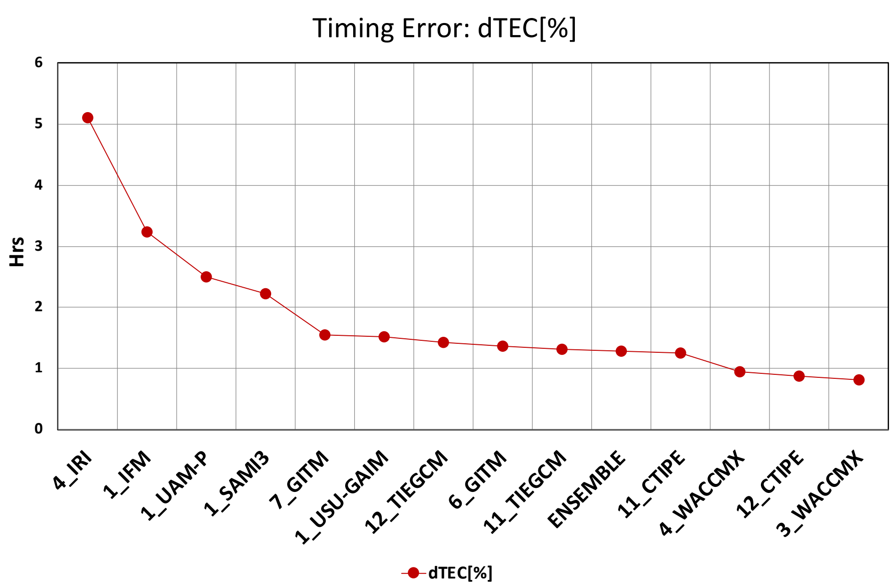

Timing Error: dTEC[%]
Hrs
6
5
4
3
2
1
0
4_IRI
1_IFM
1_UAM-P
1_SAMI3
7_GITM
1_USU-GAIM
12_TIEGCM
6_GITM
11_TIEGCM
ENSEMBLE
11_CTIPE
4_WACCMX
12_CTIPE
3_WACCMX
dTEC[%]

Figure 8

Figure 1. Scatter plots of the observed (*x* axis) and modeled (*y* axis) shifted foF2 and TEC (foF2* in the 1st , TEC* in the 3rd columns), and percentage change of foF2 and TEC (dfoF2[%] in the 2nd, dTEC[%] in the 4th columns) during the storm (03/17/2013) for all 12 locations grouped into North America (NA, green), Europe (EU, blue), South Africa (SAF, red), and South America (SAM, black)

Figure 2. Correlation Coefficients (CC) between modeled and observed foF2 (left panel) and TEC (right panel). Four CCs are displayed for each simulation: CC averaged over Europe (EU), North America (NA), Southern Hemisphere (SH refers to SAF and SAM combined), and all 12 locations, from left to right. Different colors denote different quantities. Blue denotes shifted foF2 and TEC, green and red the change and percentage changes, and black normalized percentage change. The closer the circles are to the horizontal line of 1, the better the model performances are.

Figure 3. Same as Figure 2 but for RMSE of shifted foF2 and TEC, and changes of foF2 and TEC

Figure 4. Same as Figure 2 but for RMSE of percentage change of foF2 and TEC, and normalized percentage change. Blue denotes dfoF2[%], red and black dTEC[%] and dTEC[%]_norm.

Figure 5. Same as Figure 2 but for Yield (ratio) and absolute of Timing Error (|TE| = |t_peak_model – t_peak_obs|)

Figure 6. Same as Figure 1 but for only TEC and dTEC[%] from the ensemble of the simulations (ENSEMBLE) and 1_USU-GAIM

Figure 7. Averaged CC (a) and RMSE (b) over all 12 locations of 13 simulations, the ensemble of them (ENSEMBLE), and the ensemble of 12 simulations excluding 1_USU-GAIM (ENSEMBLE_wo_DA). Blue denotes shifted TEC, green and red the change and percentage changes of TEC. CCs are plotted from the smallest to the largest (closer to 1) according to the average of the three averaged CC values of TEC, dTEC and dTEC[%]. RMSEs are plotted from the largest to the smallest according to the average RMSE for TEC and dTEC.

Figure 8. Yield and Timing Error of dTEC[%] for all 13 simulations and ENSEMBLE.

Table 1. Quantities and Skill Scores for Model-Data Comparison

Quantities and skill scores for model-data comparison	
Quiet time references	30-day median value at a given time: TEC_quiet(UT), 30 days consist of 15 days before (03/01-03/15/2013) and 15 days after (03/22-04/05/2013) the storm
Shifted TEC/foF2:	e.g., TEC*(doy, UT) = TEC(doy, UT) – minimum of TEC_quiet(UT)
TEC/foF2 changes w.r.t. the quiet time	e.g, dTEC(doy, UT)= TEC(doy, UT) –TEC_quiet (UT)
TEC/foF2 percentage changes w.r.t.the quiet time	e.g., dTEC[%](doy,UT) =100* dTEC(doy, UT)/TEC_quiet(UT)
Normalized Percentage changes of TEC	dTEC[%]_norm = (dTEC[%] -ave_dTEC[%])/std_dTEC[%]; ave_dTEC[%] is the average of dTEC[%] at a given time and at a given location over the quiet 30 days, std_dTEC[%] is the standard deviation of the average percentage change
Skill Scores	
CC	Correlation Coefficient
RMSE	Root-Mean-Square Error (=$\sqrt{\frac{\sum(x_{obs}-x_{mod})^2}{N}}$), where x_{obs} and x_{mod} are observed and modeled values
Yield	ratio of the peak of modeled percentage change to that of the observed one (=$\frac{(x_{mod})max}{(x_{obs})max}$)
Timing Error (TE)	difference between the modeled peak time and observed peak time: TE = t_peak_model – t_peak_obs

Table 2. Models used for this study

Model Setting ID	Model Version	Drivers			Upper boundary for TEC calculation/ Resolution
		Input data	Models used for thermosphere, tides from lower boundary, and high latitude electrodynamics		
Physics-based Coupled Ionosphere-Thermosphere Model					
			Tides	High Latitude Electrodynamics	
11_CTIPE[a]	CTIPe3.2 [*Codrescu et al*., 2000; *Millward et al*., 2001]	F10.7, ACE IMF data and solar wind speed and density, NOAA POES Hemispheric Power data	(2,2), (2,3), (2,4), (2,5), and (1,1) propagating tidal modes	Weimer-2005 high latitude electric potential [*Weimer*, 2005], Fuller-Rowell and Evans auroral precipitation [1987]	~2,000 km, 2° lat. × 18° long.
12_CTIPE[a]	CTIPe4.1		WAM [*Akmaev et al.,* 2011, *Fuller-Rowell et al.,* 2010] tides		
6_GITM[a]	GITM2.5 [*Ridley et al.*, 2006]	FISM solar EUV irradiance, ACE IMF data and solar wind speed and density	MSIS [*Hedin,* 1991] migrating diurnal and semidiurnal tides	Weimer-2005 high latitude electric potential, Ovation auroral precipitation [*Newell et al.,* 2009; 2011]	~600 km, 2.5° lat. × 5° long.
7_GITM	GITM21.11			Weimer-2005 high latitude electric potential, Fang's auroral precipitation [*Fang et al.,* 2013]	
12_TIE-GCM[a]	TIE-GCM2.0 [*Roble et al.*, 1988; *Richmond et al.*, 1992; *Solomon et al.*, 2012]	F10.7, Kp, OMNI IMF data and solar wind speed and density	GSWM [*Hagan et al*., 1999] migrating diurnal and semidiurnal tides	Weimer-2005 high latitude electric potential, Roble and Ridley auroral precipitation [1987]	~600 km, 2.5° lat. × 2.5° long.
Whole Atmosphere Model					
3_WACCM-X	CESM2.2 [*Gettelman et al.,* 2019; *Liu et al.*, 2018]	F10.7, Kp, OMNI IMF data and solar wind speed and density	Heelis high latitude electric potential [*Heelis et al.,* 1982], Roble and Ridley auroral precipitation [1987]		~600 km, 1.9° lat. × 2.5° long.
4_WACCM-X			Weimer-2005 high latitude electric potential, Roble and Ridley auroral precipitation [1987]		

[a]The model results are submitted by the CCMC using the models hosted at the CCMC

Table 3. Number of locations where the models correctly predict negative or positive phase.

	Time Interval	11_CTIPE	12_CTIPE	6_GITM	7_GITM	12_TIE-GCM	3_WACCM-X	4_WACCM-X
dfoF2[%]	06–15UT	**8**	7	5	**9**	9	6	**10**
	15–22UT	**10**	6	7	**8**	7	7	**10**
dTEC[%]	06–15UT	9	**10**	**10**	**10**	7	**10**	9
	15–22UT	7	**10**	**12**	11	**10**	7	8

Table 4. Averaged RMSE over all 12 locations of the ensemble of newer versions (ENSEMBLE_new) of models (12_CTIPE, 7_GITM and 4_WACCM-X) driven by Weimer2005 electric potential model, the ensemble of older versions (ENSEMBLE_old) of models (11_CTIPE, 6_GITM and 12_TIE-GCM), and 1_USU-GAIM.

	TEC (TECU)	dTEC (TECU)	dTEC[%]
ENSEMBLE_old	6.6	4.1	33.4
ENSEMBLE_new	4.6	3.2	29.8
1_USU-GAIM	4.5	3.4	29.9

Validation of Ionospheric Specifications During Geomagnetic Storms: TEC and foF2 during the 2013 March Storm Event-II

J. S. Shim[1], I.-S. Song[1], G. Jee[2], Y.-S. Kwak[3], I. Tsagouri[4], L. Goncharenko[5], J. McInerney[6], A. Vitt[6], L. Rastaetter[7], J. Yue[7,8], M. Chou[7,8], M. Codrescu[9], A. J. Coster[5], M. Fedrizzi[9], T. J. Fuller-Rowell[9], A. J. Ridley[10], S. C. Solomon[6]

[1]Department of Atmospheric Sciences, Yonsei University, Seoul, South Korea,

[2]Division of Atmospheric Sciences, Korea Polar Research Institute, Incheon, South Korea

[3]Space Science Division, Korea Astronomy and Space Science Institute, Daejeon, South Korea

[4]National Observatory of Athens, Penteli, Greece,

[5]Haystack Observatory, Westford, MA, USA,

[6]High Altitude Observatory, NCAR, Boulder, CO, USA,

[7]NASA GSFC, Greenbelt, MD, USA,

[8]Catholic University of America, Washington, DC, USA,

[9]NOAA SWPC, Boulder, CO, USA,

[10]Space Physics Research Laboratory, Univ. of Michigan, Ann Arbor, MI, USA

Corresponding author: In-Sun Song (songi@yonsei.ac.kr)

Key Points:

- foF2/TEC and their changes during a storm predicted by seven ionosphere-thermosphere coupled models are evaluated against GIRO foF2 and GPS TEC measurements.
- Model simulations tend to underestimate the storm-time enhancements of foF2 and TEC and to predict them better in the North America but worse in the southern hemisphere.
- Ensemble of all simulations for TEC is comparable to the data assimilation model (USU-GAIM).

Abstract

Assessing space weather modeling capability is a key element in improving existing models and developing new ones. In order to track improvement of the models and investigate impacts of forcing, from the lower atmosphere below and from the magnetosphere above, on the performance of ionosphere-thermosphere models, we expand our previous assessment for 2013 March storm event [*Shim et al.*, 2018]. In this study, we evaluate new simulations from upgraded models (Coupled Thermosphere Ionosphere Plasmasphere Electrodynamics (CTIPe) model version 4.1 and Global Ionosphere Thermosphere Model (GITM) version 21.11) and from NCAR Whole Atmosphere Community Climate Model with thermosphere and ionosphere extension (WACCM-X) version 2.2 including 8 simulations in the previous study. A simulation of NCAR Thermosphere-Ionosphere-Electrodynamics General Circulation Model version 2 (TIE-GCM 2) is also included for comparison with WACCM-X. TEC and foF2 changes from quiet-time background are considered to evaluate the model performance on the storm impacts. For evaluation, we employ 4 skill scores: Correlation coefficient (CC), root-mean square error (RMSE), ratio of the modeled to observed maximum percentage changes (Yield), and timing error(TE). It is found that the models tend to underestimate the storm-time enhancements of foF2 (F2-layer critical frequency) and TEC (Total Electron Content) and to predict foF2 and/or TEC better in the North America but worse in the Southern Hemisphere. The ensemble simulation for TEC is comparable to results from a data assimilation model (Utah State University-Global Assimilation of Ionospheric Measurement (USU-GAIM)) with differences in skill score less than 3% and 6% for CC and RMSE, respectively.

Plain Language Summary

The Earth's ionosphere-thermosphere (IT) system, which is present between the lower atmosphere and the magnetosphere, is highly variable due to external forcings from below and above as well as internal forcings mainly associated with ion-neutral coupling processes. The variabilities of the IT system can adversely affect our daily lives, therefore, there is a need for both accurate and reliable weather forecasts to mitigate harmful effects of space weather events. In order to track the improvement of predictive capabilities of space weather models for the IT system, and to investigate the impacts of the forcings on the performance of IT models, we evaluate new simulations from upgraded models (CTIPe model version 4.1 and GITM version 21.11) and from NCAR WACCM-X version 2.2 together with 8 simulations in the previous study. A simulation of NCAR TIE-GCM version 2 is also included for the comparison with WACCM-X. Quantitative evaluation is performed by using 4 skill scores including Correlation coefficient (CC), root-mean square error (RMSE), ratio of the modeled to observed maximum percentage changes (Yield), and timing error (TE). The findings of this study will provide a baseline for future validation studies of new and improved models.

1. Introduction

Variabilities of the Earth's ionosphere-thermosphere (IT) system, caused by charged particles and electromagnetic radiation emitted from the sun, can adversely affect our daily lives, which are highly dependent on space-based technological infrastructures such as Low-Earth Orbit (LEO) satellites and the Global Navigation Satellite System (GNSS). To mitigate harmful effects of space weather events, modeling plays a critical role in our quest to understand the connection between solar eruptive phenomena and their impacts in interplanetary space and near-Earth space environment. In particular, the Earth's upper atmosphere including the IT system is the space environment closest to the human society. Thus, during the past few decades, first-principles physics-based (PB) IT models have been developed for specifications and forecasts of the near-Earth space environment. In addition, there have been recent developments of whole atmosphere models with thermospheric and ionospheric extension to fully understand variabilities of the IT system by considering coupling between the IT system and the lower atmosphere [e.g., *Akmaev,* 2011; *Fuller-Rowell et al.,* 2010; *Jin et al*., 2011; *Liu et al*., 2018].

For more accurate space weather forecasting, assessing space weather modeling capability is a key element to improve existing models and to develop new models. Over the last decade, in an effort to address the needs and challenges of the assessment of our current knowledge about space weather effects on the IT system and current state of IT modeling capabilities, the NASA GSFC Community Coordinated Modeling Center (CCMC) has been supporting community-wide model validation projects, including Coupling, Energetics and Dynamics of Atmospheric Regions (CEDAR) [*Shim et al.*, 2011, 2012, 2014] and Geospace Environment Modeling (GEM)-CEDAR modeling challenges [*Rastätter et al*., 2016; *Shim et al.*, 2017a].

Furthermore, in 2018, the CCMC established an international effort, "International Forum for Space Weather Modeling Capabilities Assessment", to evaluate and assess the predictive capabilities of space weather models (https://ccmc.gsfc.nasa.gov/iswat/IFSWCA/). As a result of this international effort, four ionosphere/thermosphere working groups were established with an overarching goal to devise a standardized quantitative validation procedure for IT models [*Scherliess et al*., 2019].

The working group, focusing on neutral density and orbit determination at LEO, reported their initial results for specific metrics for thermosphere model assessment over the selected three full years and two geomagnetic storms in 2005 [*Bruinsma et al.*, 2018]. They reported that the tested models in general performed reasonably well, although seasonal errors were sometimes observed and impulsive geomagnetic events remain a challenge. Kalafatoglu Eyigüler et al. (2019) compared the neutral density estimates from two empirical and three PB models with those obtained from the CHAMP satellite. They suggested that several metrics that provide different aspects of the errors should be considered together for a proper performance evaluation.

Another working group, "Ionosphere Plasmasphere Density Working Team", performed the assessment of present modeling capabilities in predicting the ionospheric climatology of f_oF_2 and hmF2 for the entire year 2012 [*Tsagouri et al.,* 2018]. Tsagouri et al. (2018) identified a strong seasonal and local time dependence of the model performances, especially for PB models, which could provide useful insight for future model improvements. Tsagouri et al. cautioned that the quality of the ground truth data may play a key role in testing the model performance. Shim et al. (2018) assessed how well the ionospheric models predict storm time f_oF_2 and TEC by considering quantities, such as TEC and f_oF_2 changes and percentage changes compared to quiet time background, at 12 selected midlatitude locations in the American and European-African longitude sectors. They found that the performance of the model varies with locations, even within a localized region like Europe, as well as with the metrics considered.

In this paper, we expand our previous assessment of modeled foF2 and TEC during 2013 March storm event (17 March, 2013) [*Shim et al*., 2018] to track improvement of the models and to investigate impacts of forcings from the lower atmosphere below and from the magnetosphere above on the performance of IT models. For this study, we evaluate the updated version of the coupled IT models available at the CCMC [*Webb et al*., 2009] since our previous study [*Shim et al*., 2018]: CTIPe version 4.1 and GITM version 21.11. However, the other types of models such as empirical models, stand-alone ionospheric models, and data assimilation models are not included. In addition, for the first time, simulations of NCAR WACCM-X 2.2 are included in our assessment. We also included a simulation of NCAR TIE-GCM 2 to compare with results from WACCM-X 2.2. For TEC prediction, we compare a weighted mean of the ensemble of all 13 simulations (ensemble average), including 8 simulations from our previous study with individual simulations to assess ensemble forecast capability. In Section 2, we briefly describe observations, models, and metrics used for this study. Section 3 presents the results of model-data comparisons and performance of the models are presented. Section 4 shows comparisons of ensemble of TEC predictions with the individual simulations based on the skill scores used in this study. Finally, we summarize and conclude in Section 5.

2. Methodology

2.1 Observations and Metrics

We use the foF2 and TEC measurements at 12 ionosonde stations selected in middle latitudes: 8 northern hemisphere (NH) stations in the US (Millstone Hill, Idaho national Lab, Boulder, and Eglin AFB) and Europe (Chilton, Pruhonice, Ebre, and Athens) and 4 southern hemisphere (SH) stations in South America (Port Stanley) and South Africa (Louisvale, Hermanus, and Grahamstown) (Figure 1 and Table 1 in *Shim et al.* [2018] for details). The foF2 and GNSS vertical TEC (vTEC) data are provided by Global Ionosphere Radio Observatory (GIRO) (http://giro.uml.edu/) [*Reinisch and Galkin*, 2011] and by MIT Haystack Observatory (http://cedar.openmadrigal.org/, http://cedar.openmadrigal.org/cgi-bin/gSimpleUIAccessData.py) [*Rideout and Coster*, 2006], respectively.

Table 1 shows the quantities and skill scores calculated for model-data comparison. To remove potential systematic uncertainties in the models and observations and baseline differences among the models and between models and observations, we use the shifted values and changes from their own quiet-time background values (e.g., shifted TEC (TEC*) = TEC (UT) on a particular DOY – median (UT) of TEC for 30 days centered on the storm date). Furthermore, using these quantities likely reduce the impacts of differing upper boundaries for TEC calculations, since the plasmaspheric TEC variations with geomagnetic activity are negligible in middle latitudes [*Shim et al*., 2017b].

To measure how well the observed and modeled values are linearly correlated (in phase) with each other and how different the values are on average over the time interval considered, CC and RMSE are calculated, respectively, for the error values below 95th percentile. We also calculate Yield and timing error to measure the models' capability to capture peak disturbances during the storm. For more detailed information on the quantities and skill scores used for the study, refer to Section 2 in *Shim et al.* [2018].

2.2 Models and Simulations

The simulations used in this study are obtained from the updated and newly incorporated coupled ionosphere-thermosphere models available at the CCMC [*Webb et al*., 2009] since our previous study [*Shim et al*., 2018]: CTIPe 4.1, GITM 21.11 and WACCM-X 2.2. The WACCM-X 2.2 simulations are provided by NCAR HAO. The WACCM-X version 2 [*Liu et al.*, 2018] is a comprehensive numerical model that extends the atmospheric component model of the NCAR Community Earth System Model (CESM) [*Hurrell et al*., 2013] into the thermosphere up to 500–700 km altitude. WACCM-X is uniquely capable of being run in a configuration where the atmosphere is coupled to active or prescribed ocean, sea ice, and land components, enabling studies of thermospheric and ionospheric weather and climate. WACCM-X version 2 is based upon WACCM version 6 [*Gettelman et al.*, 2019] with a top boundary of ~130 km, which is built upon the Community Atmosphere Model (CAM) version 6 having a top boundary of ~40 km. WACCM-X 2.2 includes WACCM6 physics for middle atmosphere and lower thermosphere as well as CAM6 physics for the troposphere and the lower stratosphere, and it fully incorporates the electrodynamical processes related to low-to mid-latitude wind dynamo that is implemented in the NCAR TIE-GCM. For this study, two specified-dynamics (SD) WACCM-X 2.2 simulations with different high-latitude electrostatic potential models [*Heelis et al.*, 1982; *Weimer*, 2005] are used. The SD simulations are carried out by constraining the model's lower atmospheric neutral dynamics using meteorological reanalysis data. The constraining process is achieved by nudging the model towards MERRA-2 (Modern Era Retrospective Analysis for Research and Applications, Version 2) data [*Gelaro et al.,* 2017] below around the altitude of 50 km in a way presented by *Brakebusch et al*. [2013].

The resulting WACCM-X simulations are compared with the simulations of TIE-GCM. The comparisons between WACCM-X and TIE-GCM simulations will show differences and similarities in modeling capabilities between whole atmosphere modeling and ionosphere-thermosphere modeling with a specified low-boundary forcing (e.g., Global Scale Wave Model (GSWM) [*Hagan et al.,* 1999] used for this study).

Table 2 shows the version of the models, input data used for the simulations, and models used for lower boundary forcing and high latitude electrodynamics. We utilized unique model setting identifiers to distinguish the current simulations from those used in our previous studies [*Shim et al.*, 2011, 2012, 2014, 2017a, 2018]. Additional information for the models and model setting identifiers is available in *Shim et al*. [2011] (Refer to all references therein) and at https://ccmc.gsfc.nasa.gov/support/GEM_metrics_08/tags_list.php

To investigate improvement in foF2 and TEC predictions of the updated versions of CTIPE (12_CTIPE) and GITM (7_GITM), the simulations of the old versions of the models (11_CTIPE and 6_GITM) from our previous study are included. The comparison will be focused on the comparison between the simulations obtained from the same model. As for TIE-GCM, 12_TIE-GCM (run at 2.5° resolution) is presented for this study, but the comparison between 11_TIE_GCM and 12_TIE-GCM was not included in this study because the only difference between the two is horizontal resolution (5°lat.×5°long. vs 2.5°lat.×2.5°long.).

We should take note of the difference between the simulations obtained from the same model that influence foF2 and TEC responses to geomagnetic storms. For two CTIPe runs, different lower atmospheric tides were specified: 11_CTIPE was driven by the imposed migrating semidiurnal (2,2), (2,3), (2,4), (2,5), and diurnal (1,1) tidal modes, while 12_CTIPE was run with monthly mean spectrum of tides obtained from WAM (Whole Atmosphere Model) [*Akmaev et al.,* 2011, *Fuller-Rowell et al.,* 2010]. For two GITM simulations, 7_GITM used Fang's auroral precipitation [*Fang et al.,* 2013], while 6_GITM used Ovation model [*Newell et al.,* 2009; 2011]. For two WACCM-X simulations, Heelis and Weimer2005 electric potential models were used for 3_WACCM-X and 4_WACCM-X, respectively. 12_TIEGCM was driven by Weimer2005 electric potential model and GSWM.

3. Performance of the Models in Predictions of foF2 and vTEC on 17 March 2013

Most simulations newly added for this study show similar behavior to those used in *Shim et al.* [2018], in predicting foF2 and TEC during the storm. For example, the simulations are not able to reproduce (1) the difference between eastern and western parts of the North American sector (e.g., TEC increases at Millstone Hill but decreases at Idaho and Boulder around 20UT), and (2) different responses between foF2 (negligible changes) and TEC (noticeable increase) found in European (Chilton) and South-African (Grahamstown) stations (See Figure 4 of Shim et al. [2018] for reference). However, compared to other simulations, 4_WACCM-X driven by *Weimer* (2005) high latitude electric potential model captures relatively well the two differences in TEC and foF2 described above (Figure S1 in supporting information).

Figure 1 shows scatter plots of the observed (*x* axis) and modeled (*y* axis) shifted foF2 and TEC, and percentage change of foF2 and TEC during the storm (03/17/2013) for all 12 locations grouped into 4 sectors: North America (NA, green), Europe (EU, blue), South Africa (SAF, red), and South America (SAM, black). First of all, the qualitative comparison between the simulations from the same model can be summarized as follows. 11_CTIPE/12_CTIPE tends to underestimate foF2 for both quiet and disturbed conditions, but 12_CTIPE predicts much better both foF2 and TEC during the storm than 11_CTIPE. 6_GITM and 7_GITM underestimate foF2 and TEC for all cases and show relatively small response to the storm compared to the other simulations. 12_TIE-GCM and WACCM-Xs produce similar foF2 and TEC changes during the storm. All three simulations give *substantial underestimation of TEC in SAF.* 12_TIE-GCM and *3_WACCM-X* produce larger overestimation of foF2 and TEC in NA sector than *4_WACCM-X. 4_WACCM-X* shows substantial improvement in the TEC overestimation in NA. 3_WACCM-X, of which the high latitude electric potential is specified by Heelis et al. [1982], tends to overestimate foF2 and TEC compared with 4_WACCM-X. 3_WACCM-X and 4_WACCM-X produce better quiet time foF2 and TEC than 12_TIEGCM does and capture wave-like small increases in foF2 and TEC at Idaho National Lab around 10–11UT (2–3 LT) (Figure S1 in supporting information).

As shown for 6_GITM and 11_CTIPE in *Shim et al.* [2018], the modeled foF2 values of 7_GITM and 12_CTIPE better agrees with the observed ones when they are shifted by subtracting the minimum of 30-day median (see Figure S2 in supporting information, *Shim et al.* [2018]). Most foF2 and TEC data points of 7_GITM and 12_CTIPE before shifting are below and above the line with slope 1 (black solid line), respectively. This indicates that 7_GITM underestimates foF2 and TEC like 6_GITM, while 12_CTIPE overestimates them. The models that tend to underestimate foF2, such as 6_GITM, 7_GITM and 11_CTIPE, seem to unable to produce foF2* larger than about 7 MHz, and underestimate TEC* being less than about 20 TECU during the storm as reported in *Shim et al.* [2018]. 12_TIE-GCM and WACCM-Xs show similar distribution of the data points after shifting foF2 and TEC with a tendency to underestimate foF2 and TEC in the South Africa region.

The modeled dfoF2[%] and dTEC[%] show less agreement with the observed values than the modeled foF2* and TEC* do. The data points in the 2nd quadrant (top left) and the 4th quadrant (bottom right) indicate that the modeled and observed percentage changes are in opposite sign. 7_GITM and 3_WACCM-X have more data points in the 2nd quadrant for dfoF2[%] prediction than 6_GITM and 4_WACCM-X, respectively. Like most simulations used in our previous evaluation [*Shim et al.* 2018], 12_CTIPE and 7_GITM do not appear to reproduce the large dTEC[%] (about 200 %) at Port Stanley in SAM. However, 12_TIE-GCM and WACCM-Xs better produce the enhancement in TEC percentage change. Compared to 4_WACCM-X and 12_TIE-GCM, 3_WACCM-X overestimates dTEC[%] especially in NA and EU regions. 12_CTIPE and 6_GITM have more data points of overestimated dTEC[%] in SAF than 11_CTIPE and 7_GITM, respectively.

From now on, foF2 and TEC will represent shifted foF2 (foF2*) and shifted TEC (TEC*), respectively.

3.1 Correlation Coefficient (CC)

We first calculate correlation coefficient (CC) between the modeled and observed foF2 and TEC for DOY 076 (17 March, 2013) for quantitative assessment of the model performance of TEC and foF2 predictions. In Figure 2, the CCs for each simulation are presented for foF2 in the left panel and for TEC in the right panel. For each simulation, four CC values are displayed. First three of the values correspond to the average CC over Europe (EU), North America (NA), Southern Hemisphere (SH refers to SAF and SAM combined), and the last one is the average of all 12 locations. The modeled foF2 and TEC (blue dots) are highly correlated with the observed values. The average CC values over all 12 locations for both foF2 and TEC are about 0.8–0.95, but the average CCs for their changes are much smaller. For example, the CCs for TEC changes (dTEC) are 0.5–0.6 and even smaller for foF2. The modeled foF2 changes (green), percentage changes (red) and normalized percentage changes (black only applicable for TEC) are much less correlated (closer to uncorrelated) with the observed values (about $0.1 < \text{average CC} < 0.4$). There is no big difference between dTEC[%] and dTEC[%]_norm based on the average values for each simulation as reported in *Shim et al.* [2018].

Note that the CC values for the changes and percentage changes of foF2 and TEC are highly dependent on locations. Most simulations, except for 12_CTIPE and GITMs, show lower CC for dfoF2 and dTEC in NA. It seems to be caused by the decreases of foF2 and TEC during the storm (negative phase) in the western parts of NA that are not captured well. GITMs show the negative phase well although it underestimated the magnitude of the change. The CCs for the percentage changes of foF2 and TEC are particularly small for CTIPEs and GITMs.

11_CTIPE's foF2 and TEC averaged over 12 locations are slightly better correlated with the observed values than 12_CTIPE. However, the changes and percentage changes of foF2 and TEC from 12_CTIPE are better correlated with the observed values than 11_CTIPE's values in most regions. Although the two GITMs produce similar CCs, 7_GITM shows better CC in NA regions for dfoF2, dfoF2[%], dTEC[%], and n_dTEC[%], while 6_GITM shows better CC for foF2 and dTEC. WACCM-Xs perform better than 12_TIE_GCM for all the considered quantities based on the average except for dTEC. WACCM-Xs perform similar to each other.

Close inspection of Figures. 1 and 2 indicates that a linearity between CTIPE and observations is improved in the newer version of CTIPE (12_CTIPE), but 12_CTIPE gives more scattered distribution around a linear relation (Fig. 1), which seems to lead to the lower CC in 12_CTIPE than in 11_CTIPE. 7_GITM exhibits a slight improvement in a linearity between the model and observations (Fig. 1), but this improvement is not clearly seen in the correlation analysis (Fig. 2). For 12_TIEGCM and WACCM-Xs, both a linearity between the models and observations (Fig. 1) and CCs (Fig. 2) demonstrate that the model performances are overall improved in WACCM-Xs compared with TIEGCM. In terms of the model-observation linearity, 4_WACCMX is somewhat better than 3_WACCMX (Fig. 1), but their CCs seems comparable to each other (Fig. 2).

3.2 Root Mean Square Error (RMSE)

Figure 3 shows RMSE of foF2 and dfoF2 in the left panel, and TEC and dTEC in the right panel. For foF2 (blue) and dfoF2 (green) predictions, based on the average RMSE values, the

RMSEs from the updated version (12_CTIPE and 7_GITM) are about 1.5 MHz for foF2 and about 1 MHz for dfof2, and they are slightly lower than RMSEs in their old versions. 12_CTIPE shows improvement in foF2 in SH and dfoF2 in NA and EU compared to 11_CTIPE. 7_GITM performs better in foF2 and dfoF2 in EU and SH than 6_GITM. 4_WACCM-X has smaller RMSE (~1 MHz) than 3_WACCM-X and 12_TIE-GCM (~1.3 MHz for dfoF2 and ~2 MHz for foF2).

12_CTIPE is better in TEC prediction than 11_CTIPE, while the opposite holds true for dTEC prediction. The two GITMs' average RMSE values for TEC and dTEC predictions are similar to each other, about 9 TECU for TEC and 5 TECU for dTEC. Like foF2 and dfoF2 prediction, 4_WACCM-X has smaller RMSE (~ 5 TECU for TEC and 4 TECU for dTEC) than 12_TIE-GCM and 3_WACCM-X (~6 TECU).

As seen in *Shim et al.* [2018], RMSE is highly variable with location. Most simulations appear to predict foF2 and/or TEC better in NA and worse in SH (except for 12_TIE-GCM for foF2 and 12_CTIPE for TEC). Both 11_CTIPE and GITMs tend to perform better in NA for dTEC, while WACCM-Xs show the opposite tendency for dfoF2 and dTEC. 7_GITM and 4_WACCM-X shows the least RMSE dependence on location for dfoF2 and for dTEC, respectively, among seven simulations.

Figure 4 shows the RMSE of percentage changes of foF2 (blue) and TEC (red) and normalized percentage changes of TEC (black). The two CTIPEs produce the similar RMSE for dTEC[%], but 12_CTIPE and 11_CTIPE produce lower RMSE for dfoF2[%] and dTEC[%]_norm, respectively. For all three percentage changes of dfoF2[%], dTEC[%], and dTEC[%]_norm, 7_GITM seems to perform better than 6_GITM based on the average RMSEs over the 12 locations. 4_WACCM-X and 12_TIE-GCM perform very similarly for dfoF2[%] and dTEC[%] and better than 3_WACCM-X.

Difference in the performance among locations is more noticeable in dTEC[%] and dTEC[%]_norm than in dfoF2[%] as found in *Shim et al.* [2018]. All simulations, except 6_GITM, produce lower RMSE of dTEC[%] in NA and higher in SH region. This tendency remains the same for dTEC[%]_norm with the exception of 3_WACCM-X, which has lower RMSE for dTEC[%]_norm in SH. For 3_WACCM-X, the higher RMSE for dTEC[%] and the lower RMSE for dTEC[%]_norm in SH than in NA are probably due to the normalization factor, standard deviation of dTEC[%] in the locations.

3.3 Yield and Timing Error (TE)

To measure how well the models capture the degree of TEC and foF2 disturbances during the main phase, Yield and Timing Error (TE) of dfoF2[%], dTEC[%], and dTEC[%]_norm are calculated. *Shim et al.* [2018] considered two time intervals, 06–15UT and 15–22UT, when peaks are observed in most of 12 locations. In each time interval, we calculate one Yield value and one TE value. Definitions of Yield and TE are presented in Table 1.

In each sector, average Yield and TE are calculated over the number of stations where the model correctly predicts the storm phase, i.e., Yield is positive. Table 3 shows the total number of stations where the models show correct storm phase, either positive or negative. The numbers in bold are the higher values between the simulations compared. 12_CTIPE predicts the storm phase better for dTEC[%] than 11_CTIPE, but 11_CTIPE predicts better for dfoF2[%] than 12_CTIPE. 7_GITM is improved in predicting the storm phase of dfoF2[%], while 6_GITM predicts better the storm phase of dTEC[%]. 4_WACCM-X, compared to 12_TIE-GCM and

3_WACCM-X, is better for predicting the phase of dfoF2[%] and worse for predicting that of dTEC[%].

Figure 5 shows average Yield (left) and average of absolute values of TE (right) over the two time intervals: dfoF2[%] in blue, dTEC[%] in red, and dTEC[%]_norm in black. Concerning the average of all 12 locations, 12_CTIPE appears to overestimate peak values of dTEC[%] and dTEC[%]_norm with larger variation with location (e.g., ~1 < Yield of dTEC[%]_norm < ~2.5) than 11_CTIPE, of which Yield is less than 1 for all three quantities of percentage changes (e.g., 0.7 < Yield of dTEC[%]_norm < 0.9). Yields of 12_CTIPE for dTEC[%] and dTEC[%]_norm are closer to 1 in NA. GITMs produce similar ratios based on the average over all locations, but 7_GITM shows smaller differences in Yield among locations (e.g., ~0.5 < Yield of dTEC[%]_norm < ~1) than 6_GITM (e.g., 0.5 < Yield of dTEC[%]_norm < ~2.5). In terms of average Yield, 12 TIE-GCM and two WACCM-Xs tend to overestimate the peak values and show similar performance, although 12_TIE-GCM's ratios are closer to 1 than those of WACCM-Xs. 3_WACCM-X shows larger variation in Yield among locations (e.g., ~0.9 < Yield of dTEC[%]_norm < ~2.7) than 12_TIE-GCM and 4_WACCM-X (e.g., ~1.7 < Yield of dTEC[%]_norm < ~2.3).

Average Timing Errors of dfoF2[%] and dTEC[%]_norm are between 1 and 2 hours, and TE of dTEC[%] are about 0.8–1.5 hours. With respect to the average TE, 12_CTIPE has smaller TE (~1 hr) than 11_CTIPE (about 1.5 hr) for all three percentage changes with less location dependence as well. 7_GITM's three TEs are about 1.5 hrs, while 6_GITM's TEs of dfoF2[%], dTEC[%] and dTEC[%]_norm are ~1, ~1.4, and ~2 hrs, respectively. 12 TIE-GCM has smaller TE for dfoF2[%] and 3_WACCM-X has smaller TE for dTEC[%] and dTEC[%]_norm, however 3_WACCM-X show larger location dependence of TE for dTEC[%]_norm and dfoF2[%].

4. Ensemble of TEC obtained from13 simulations

The linearity check, RMSE, and CC between model results and observations for shifted foF2 and TEC and their relative changes indicate that the newer versions of the models (i.e., 12_CTIPE, 7_GITM and 4_WACCM-X) produces the better results. From the viewpoints of correct prediction of storm phases (Table 3), Yields, and TEs (Fig. 5), however, there is no one best simulation for all locations, and the performance of model varies with locations as well as the Yields and TE.

The differences in performance among the simulations could be caused by inherent differences among the models or by a combination of different input data and different models used for lower boundary forcing and high-latitude electrodynamics. Even different data assimilation models for the same weather condition can yield different results, due to numerous reasons (e.g., the use of different background weather models, spatial/temporal resolutions, assimilation methods, and data error analyses), even if the same data are assimilated [*Schunk et al.*, 2021]. The common way to handle these differences is to use model ensembles and the use of ensembles enables estimations of the certainty of results. Thus, we used a weighted mean of the ensemble of all 13 simulations including 8 simulations from our previous study (*Shim et al.,* 2018) for TEC, dTEC and dTEC[%] to compare the ensemble average with the individual simulations. To get the weighted mean ($\bar{x} = \sum w_i x_i / \sum w_i$), we used the RMSE of shifted TEC ($w_i = 1/\mathrm{RMSE}$).

Figure 6 is the same as Figure 1 but for the ensemble of the simulations (ENSEMBLE will be used as model setting ID) and a simulation (1_USU-GAIM) from a data assimilation model (DA), USU-GAIM. For TEC less than about 20 TECU, ENSEMBLE shows better agreement with GPS TEC than the individual simulations, including 1_USU-GAIM. However, as we can expect, ENSEMBLE underestimates TEC larger than about 30 TECU due to the tendency to underestimate TEC of many simulations as pointed out in Section 3 and *Shim et al.,* [2018]. For dTEC[%], ENSEMBLE appears to be correlated better with GPS dTEC[%] than the other simulations, although there are some underestimations in SAF, as well as in SAM with opposite prediction of the storm phase.

Figure 7 shows averaged CC and RMSE values over all 12 locations of 13 simulations, the ensemble of them, and the ensemble of 12 simulations excluding 1_USU-GAIM (ENSEMBLE_wo_DA). The simulations in Figure 7 (a) were arranged by the average of the three averaged CC values for TEC, dTEC and dTEC[%] from the smallest to the largest (closer to 1). In Figure 7 (b), the simulations were arranged by the average of the two averaged RMSEs for TEC and dTEC from the largest to the smallest. Based on the averaged CC and RMSE, ENSEMBLEs (ENSEMBLE and ENSEMBLE_wo_DA) of the simulations perform very similarly and outperform all 12 simulations but a data assimilation model, 1_USU-GAIM. However, ENSEMBLEs and 1_USU-GAIM do not show big difference in their performance. The differences in RMSE of TEC and dTEC between ENSEMBLE and 1_USU-GAIM are less than 0.5 and 0.1 TECU, respectively. For dTEC[%], ENSEMBLE performs slightly better than 1_USU-GAIM with about 1.5% lower RMSE. The fact that ENSEMBLEs are comparable to the data assimilation model 1_USU-GAIM indicates that the multi-model ensemble can be useful in forecasting the IT system, although this result is obtained from a single geomagnetic storm event.

Figure 8 shows Yield and Timing Error of dTEC[%] for all 13 simulations along with ENSEMBLE. The values correspond to the average over all 12 locations. Unlike CC and RMSE, ENSEMBLE does not outperform all physic-based coupled models in terms of Yield and TE, although the difference is small. ENSEMBLE underestimates Yield, while most of the simulations overestimate it, except 4_IRI and 11_CTIPE. 7 simulations from PB coupled IT models and 1_USU-GAIM produce Yield closer to 1 than ENSEMBLE does.

Timing Error of dTEC[%] of ENSEMBLE is about 1 hr, which is slightly larger than TE from 4 simulations from CTIPE and WACCM-X, but the difference from the smallest TE is less than 0.5 hr.

Regarding the averaged skill scores for all 12 locations, newly added five simulations in this study produce comparable TEC and TEC changes to the simulations from PB IT models used in our previous study. The simulations of newer versions of the models (12_CTIPE, 7_GITM and 4_WACCM-X) are found to give overall improved forecast results. Based on the averaged RMSE, the ensemble of simulations of the models' newer versions is comparable to 1_USU-GAIM and performs better than the ensemble of the simulations of old versions of models (11_CTIPE, 6_GITM and 12_TIE-GCM) (Table 4).

5. Summary and Conclusions

We expanded on our previous systematic assessment of modeled foF2 and TEC during 2013 March storm event (17 March, 2013) to track the improvement of the models and investigate impacts of forcings from the lower atmosphere and the magnetosphere, on the performance of ionosphere-thermosphere coupled models.

We evaluated simulations from upgraded models (CTIPe4.1 and GITM21.11) since our previous assessment and a whole atmosphere model (WACCM-X2.2). To compare with results from WACCM-X2.2, we also included a simulation of TIE-GCM2.0, of which the electrodynamic processes are implemented in WACCM-X 2.2. Furthermore, to evaluate TEC prediction of the simulations, we used a weighted mean of the ensemble of all 13 simulations including 8 simulations from our previous study to compare the ensemble average with the individual simulations.

For evaluation of the simulations, we used the exact same procedure with the same data set, same physical quantities, and same skill scores as our previous study [*Shim et al.*, 2018]. The skill scores were calculated for the three sectors, EU (Europe), NA (North America), and SH (Southern Hemisphere) to investigate the longitudinal and hemispheric dependence of the performance of the models.

From the five simulations used in the study, we also found the general behaviors of most simulations identified in *Shim et al.* [2018]: 1) tendency to underestimate storm-time enhancements of foF2 and TEC and not to reproduce large enhancements of dTEC[%] (e.g., about 200 % TEC increase at Port Stanley in the SAA region), 2) being unable to capture opposite responses to the storm in the eastern and western parts of NA, especially negative phase (except for GITM), which is what in part causes lower CC in NA, 3) tendency to predict foF2 and/or TEC better in NA and worse in SH with respect to RMSE. However, it was found that 12_TIE-GCM and WACCM-Xs better produce the large TEC percentage changes at Port Stanley in SAM. Based on the averaged skill scores for all 12 locations, the five simulations used in this study show skill scores better or comparable to those of the simulations from PB IT models used in our previous study.

Compared to 11_CTIPE (obtained from CTIPe3.2), 12_CTIPE (from CTIPe4.1) driven by tides from WAM tends to overestimate foF2 and TEC for both quiet and disturbed conditions and predicts better TEC peaks during the storm. For more cases, 12_CTIPE performs largely better than 11_CTIPE based on the average scores. 12_CTIPE predicts the storm phase better for dTEC[%], but 11_CTIPE does better for dfoF2[%]. 12_CTIPE appears to overestimate peak values of dTEC[%] and dTEC[%]_norm, while 11_CTIPE produces Yield less than 1.

The two GITMs, 7_GITM with Fang's auroral precipitation and 6_GITM with Ovation model, underestimate foF2 and TEC for all cases and show relatively small response to the storm compared to the other simulations that do not appear to reproduce the large dTEC[%] (about 200 % increase at Port Stanley in SAM). 7_GITM and 6_GITM perform very similarly for most cases with similar skill scores. However, 7_GITM shows better CC for most quantities except for dTEC, and lower RMSEs and Yield closer to 1 for most regions and quantities considered. 7_GITM shows the least RMSE dependence on location for dfoF2 among the other simulations.

Comparing two WACCM-Xs and 12_TIE-GCM, the two WACCM-Xs, 3_WACCM-X with Heelis high latitude electric potential model and 4_WACCM-X with Weimer 2005, predict quiet time foF2 and TEC better than 12_TIE-GCM. During the storm, 12_TIE-GCM and 4_WACCM-X produce similar foF2 and TEC in NA sector, while 3_WACCM-X tends to overestimate them and produces larger changes in foF2 and TEC. In most cases, WACCM-Xs and 12_TIE_GCM perform similarly in terms of average values of skill scores, but 3_WACCM-X and/or 4_WACCM-X perform better than 12_TIE-GCM except for Yield of percentage changes. 4_WACCM-X slightly outperforms 3_WACCMX for all cases but not for TE for percentage changes.

Our findings suggest that the newer versions of the models (12_CTIPE, 7_GITM and 4_WACCM-X) with Weimer2005 electric potential model give overall improved forecast, and the performance of the models depends on forcing from the magnetosphere and also forcing from the lower atmosphere even during storms.

For TEC, dTEC and dTEC[%], our results indicate that the ensemble of all 13 simulations (ENSEMBLE), including 8 simulations from our previous study (*Shim et al.,* 2018) is comparable to the data assimilation model (1_USU-GAIM) with differences in skill score less than 3% and 6% for CC and RMSE, respectively. However, ENSEMBLE underestimates Yield (0.73) while 7 simulations from PB coupled IT models and 1_USU-GAIM produce Yield closer to 1. Timing Error of dTEC[%] of ENSEMBLE is about 1 hr, but the difference from the smallest TE of the simulations is less than 0.5 hr. In addition, based on RMSE, the ensemble of the newer versions of the models (12_CTIPE, 7_GITM and 4_WACCM-X) is comparable to 1_USU-GAIM.

To advance our understanding of the ionosphere-thermosphere system requires significant efforts to improve the capability of numerical models along with the scope of observations [*Heelis and Maute*, 2020]. There have been recent new developments of theoretical models, including AMGeO (Assimilative Mapping of Geospace Observations) for High-Latitude Ionospheric Electrodynamics [*Matsuo*, 2020] and MAGE geospace model that couples the Grid Agnostic MHD for Extended Research Applications (GAMERA) global MHD model of the magnetosphere (Sorathia et al., 2020; Zhang et al., 2019), the Rice Convec-tion Model (RCM) model of the ring current (Toffoletto et al., 2003), TIEGCM of the upper atmosphere and the RE-developed Magnetosphere-Ionosphere Coupler/Solver (REMIX) (Merkin & Lyon, 2010). These models will be available soon to the public through CCMC, and then the modeling capability will help us better understand the processes responsible for the observed characteristics and features during disturbed conditions. In addition, CCMC will also provide users with the capability to run PB IT models with various combination of models for lower atmospheric forcing and for magnetosphere forcing, which enable us to research further the impacts of the forcings on the IT system.

The findings of this study will provide a baseline for future validation studies using new models and improved models, along with earlier results [*Shim et al.*, 2011, 2012, 2014, 2017a, 2018] obtained through CEDAR ETI, GEM-CEDAR Modeling Challenges, and the international effort, "International Forum for Space Weather Modeling Capabilities Assessment". We will extend our study to include more geomagnetic storm events to investigate differences and similarities in the performance of the models. In addition, we will also include foF2 and TEC predictions for the high- and low-latitude regions.

Acknowledgement

This work was supported by Korea Polar Research Institute (KOPRI) grant funded by the Ministry of Oceans and Fisheries (KOPRI PE22020) and basic research funding from the Korea Astronomy and Space Science Institute (KASI) (KASI2022185009). The vertical TEC data were provided by MIT Haystack Observatory and can be obtained through CEDAR Madrigal database (http://cedar.openmadrigal.org). We thank the operators of the digisondes for sharing their data through http://giro.uml.edu/. Data from the South African Ionosonde network is made available through the South African National Space Agency (SANSA), who are acknowledged for facilitating and coordinating the continued availability of data. This work is supported by grants from the National Science Foundation (NSF) Space Weather Program. This model validation study is supported by the Community Coordinated Modeling Center (CCMC) at the Goddard Space Flight Center. Data processing and research at MIT Haystack Observatory are supported by cooperative agreement AGS-1242204 between the U.S. National Science Foundation and the

Massachusetts Institute of Technology. The National Center for Atmospheric Research is sponsored by the National Science Foundation. Model output and observational data used for the study will be permanently posted at the CCMC website (http://ccmc.gsfc.nasa.gov) and provided as a resource for the space science community to use in the future.

References

Akmaev, R. A. (2011). Whole atmosphere modeling: Connecting terrestrial and space weather. *Reviews of Geophys. 49*, RG4004. 390 https://doi.org/10.1029/2011RG000364

Brakebusch, M., Randall, C. E., Kinnison, D. E., Tilmes, S., Santee, M. L., and Manney, G. L. (2013) Evaluation of Whole Atmosphere Community Climate Model simulations of ozone during Arctic winter 2004–2005, J. Geophys. Res., 118, 2673–2688, https://doi.org/10.1002/jgrd.50226

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

Chamberlin, P. C., Woods, T. N., & Eparvier, F. G. (2007). Flare Irradiance Spectral Model (FISM): Daily component algorithms and results. *Space Weather, 5,* S07005. https://doi.org/10.1029/2007SW000316

Codrescu, M. V., T. J. Fuller-Rowell, J. C. Foster, J. M. Holt, and S. J. Cariglia, (2000), Electric field variability associated with the Millstone Hill electric field model, *J. Geophys. Res*., 105, 5265–5273, doi:10.1029/1999JA900463.

Fang, X., D. Lummerzheim, and C. H. Jackman (2013), Proton impact ionization and a fast calculation method, J. Geophys. Res. Space Physics, 118, 5369–5378, doi:10.1002/jgra.50484.

Fuller -Rowell, T. J., and D. S. Evans, (1987), Height-Integrated Pedersen and Hall Conductivity Patterns Inferred From the TIROS-NOAA Satellite Data, *J. Geophys. Res.*, 92(A7), 7606–7618.

Fuller-Rowell, T., Wu, F., Akmaev, R., Fang, T.-W., & Araujo-Pradere, E. (2010). A whole atmosphere model simulation of the impact of a sudden stratospheric warming on thermosphere dynamics and electrodynamics. *Journal of Geophysical Research*, *115*, A00G08. https://doi.org/10.1029/2010JA015524

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *Journal of Climate*, 30(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., et al.(2019). The whole atmosphere community climate model version 6 (WACCM6), *Journal of Geophysical Research: Atmospheres*, *124*, 12,380–12,403. https://doi.org/10.1029/2019JD030943.

Hagan, M. E., M. D. Burrage, J. M. Forbes, J. Hackney, W. J. Randel, and X. Zhang, (1999), GSWM-98: results for migrating solar tides. *J. Geophys. Res.* 104: 6813–6828.

Hedin, A. E. (1991), Extension of the MSIS thermospheric model into the middle and lower atmosphere, *J. Geophys. Res*., 96, 1159–1172.

Heelis, R. A., J. K. Lowell, and R. W. Spiro, (1982), A Model of the High-Latitude Ionospheric Convection Pattern, *J. Geophys. Res*. 87, 6339.

Heelis, R. A., & Maute, A. (2020). Challenges to understanding the Earth's ionosphere and thermosphere. *JGR: Space Physics*, *125*, https:// doi.org/10.1029/2019JA027497

Jin, H., Miyoshi, Y., Fujiwara, H., Shinagawa, H., Terada, K., Terada, N., et al. (2011). Vertical connection from the tropospheric activities to the ionospheric longitudinal structure simulated by a new Earth's whole atmosphere-ionosphere coupled model. *Journal of Geophysical Research*, *116*, A01316. https://doi.org/10.1029/2010JA015925

Kalafatoglu Eyiguler, E. C., Shim, J. S.,Kuznetsova, M. M., Kaymaz, Z.,Bowman, B. R., Codrescu, M. V., 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.

Liu, H.-L., Bardeen, C. G., Foster, B. T., Lauritzen, P., Liu, J., Lu, G., . . . Wang, W. (2018). Development and validation of the Whole Atmosphere Community Climate Model with thermosphere and ionosphere extension (WACCM-X 2.0), *Journal of Advances in Modeling Earth Systems*, 10. https://doi.org/10.1002/2017MS001232

Matsuo, T. (2020). Recent Progress on Inverse and Data Assimilation Procedure for High-Latitude Ionospheric Electrodynamics. In: Dunlop, M., Lühr, H. (eds) Ionospheric Multi-

Spacecraft Analysis Tools. ISSI Scientific Report Series, vol 17. Springer, Cham. https://doi.org/10.1007/978-3-030-26732-2_10

Merkin, V., & Lyon, J. (2010). Effects of the low-latitude ionospheric boundary condition on the global magnetosphere. *Journal of Geophysical Research*, *115*(A10). A10202. https://doi.org/10.1029/2010JA015461

Millward, G. H., I. C. F. Müller-Wodrag, A. D. Aylward, T. J. Fuller-Rowell, A. D. Richmond, and R. J. Moffett, (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, 24,733–24,744, doi:10.1029/2000JA000342.

Newell, P. T., T. Sotirelis, and S. Wing (2009), Diffuse, monoenergetic, and broadband aurora: The global precipitation budget, *J. Geophys. Res*., 114, A09207, doi: 10.1029/2009JA014326.

Newell, P.T., and J.W. Gjerloev (2011), Substorm and magnetosphere characteristic scales inferred from the SuperMAG auroral electrojet indices, *J. Geophys. Res*., 116, A12232, doi:10.1029/2011JA016936.

Rastäetter, L., et al., (2016), GEM-CEDAR Challenge: Poynting Flux at DMSP and modeled Joule Heat, *Space Weather*, *14*, 113–135, doi:10.1002/2015SW001238.

Reinisch, B., and I. Galkin, (2011). Global Ionospheric Radio Observatory (GIRO). *Earth, Planets, and Space. 63*. 377-381. 10.5047/eps.2011.03.001.

Richmond, A. D., E. C. Ridley and R. G. Roble, (1992), A Thermosphere/Ionosphere General Circulation Model with coupled electrodynamics, *Geophys. Res. Lett.*, **19**, 601-604.

Rideout, W., and A. Coster, (2006), Automated GPS processing for global total electron content data, GPS Solution, doi:10.1007/s10291-006-0029-5.

Ridley, A. J., Y. Deng, and G. Toth, (2006), The global ionosphere-thermosphere model, *J. Atmos. Sol. Terr. Phys*., 68, 839-864.

Roble, R. G., E. C. Ridley, A. D. Richmond, and R. E. Dickinson, (1988), A coupled thermosphere/ionosphere general circulation model, *Geophys. Res. Lett*., 15, 1325–1328, doi:10.1029/GL015i012p01325.

Scherliess, L., Tsagouri, I., Yizengaw, E., Bruinsma, S., Shim, J. S., Coster, A., and Retterer, J. M. (2019). The International Community Coordinated Modeling Center space weather modeling capabilities assessment: Overview of ionosphere/thermosphere activities. *Space Weather*, *17*. https:// doi.org/10.1029/2018SW002036

Schunk, R. W., Scherliess, L., Eccles, V., Gardner, L. C., Sojka, J. J., Zhu, L., et al. (2021). Challenges in specifying and predicting space weather. *Space Weather*, *19*, e2019SW002404. https:// doi.org/10.1029/2019SW002404

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., et al., (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*, S10004, doi:10.1029/2012SW000851.

Shim, J. S., et al., (2014), Systematic Evaluation of Ionosphere/Thermosphere (IT) Models: CEDAR Electrodynamics Thermosphere Ionosphere (ETI) Challenge (2009-2010), in *Modeling the Ionosphere-Thermosphere System*, AGU Geophysical Monograph Series.

Shim, J. S., Rastätter, L., Kuznetsova, M., Bilitza, D., Codrescu, M., Coster, A. J., … Zhu, L. (2017a). CEDAR-GEM challenge for systematic assessment of Ionosphere/thermosphere models in predicting TEC during the 2006 December storm event. *Space Weather, 15*, 1238–1256. https://doi.org/10.1002/ 2017SW001649

Shim, J. S., G. Jee, and L. Scherliess (2017b), Climatology of plasmaspheric total electron content obtained from Jason 1 satellite, J. Geophys. Res. Space Physics, 122, 1611–1623, doi:10.1002/2016JA023444.

Shim, J. S., Tsagouri, I., Goncharenko, L., Rastaetter, L., Kuznetsova, M., Bilitza, D., et al. (2018). Validation of ionospheric specifications during geomagnetic storms: TEC and foF2 during the 2013 March storm event. Space Weather, 16, 1686–1701. https://doi.org/10.1029/ 2018SW002034

Solomon, S. C., A. G. Burns, B. A. Emery, M. G. Mlynczak, L. Qian, W. Wang, D. R. Weimer, and M. Wiltberger (2012). Modeling studies of the impact of high-speed streams and co-rotating interaction regions on the thermosphere-ionosphere. *J. Geophys. Res.*, *117*, A00L11, doi:10.1029/2011JA017417

Sorathia, K., Merkin, V., Panov, E., Zhang, B., Lyon, J., Garretson, J., et al. (2020). Ballooning-interchange instability in the near-Earth plasma sheet and auroral beads: Global magnetospheric modeling at the limit of the MHD approximation. *Geophysical Research Letters*, *47*(14), e2020GL088227. https://doi.org/10.1029/2020GL088227

Tsagouri, I., Goncharenko, L., Shim, J. S., Belehaki, A., Buresova, D., & Kuznetsova, M. (2018). Assessment of current capabilities in modeling the ionospheric climatology for space weather applications: foF2 and hmF2. *Space Weather*, *16*, 1930–1945. https://doi.org/10.1029/2018SW002035

Toffoletto, F., Sazykin, S., Spiro, R., & Wolf, R. (2003). Inner magnetospheric modeling with the rice convection model. *Space Science Reviews*, *107*(1–2), 175–196. https://doi.org/10.1023/A:1025532008047

Webb, P. A., M. M. Kuznetsova, M. Hesse, L. Rastaetter, and A. Chulaki, (2009), Ionosphere-thermosphere models at the Community Coordinated Modeling Center, *Radio Sci*., 44, RS0A34, doi:10.1029/2008RS004108.

Weimer, D. R., (2005), Improved ionospheric electrodynamic models and application to calculating Joule heating rates, *J. Geophys. Res*., *110,* A05306, doi:10.1029/2004JA010884.

Zhang, B., Sorathia, K. A., Lyon, J. G., Merkin, V. G., Garretson, J. S., & Wiltberger, M. (2019). GAMERA: A three-dimensional finite-volume MHD solver for non-orthogonal curvilinear geometries. *The Astrophysical Journal Supplement Series*, *244*(1), 20. https://doi.org/10.3847/1538- 4365/ab3a4c

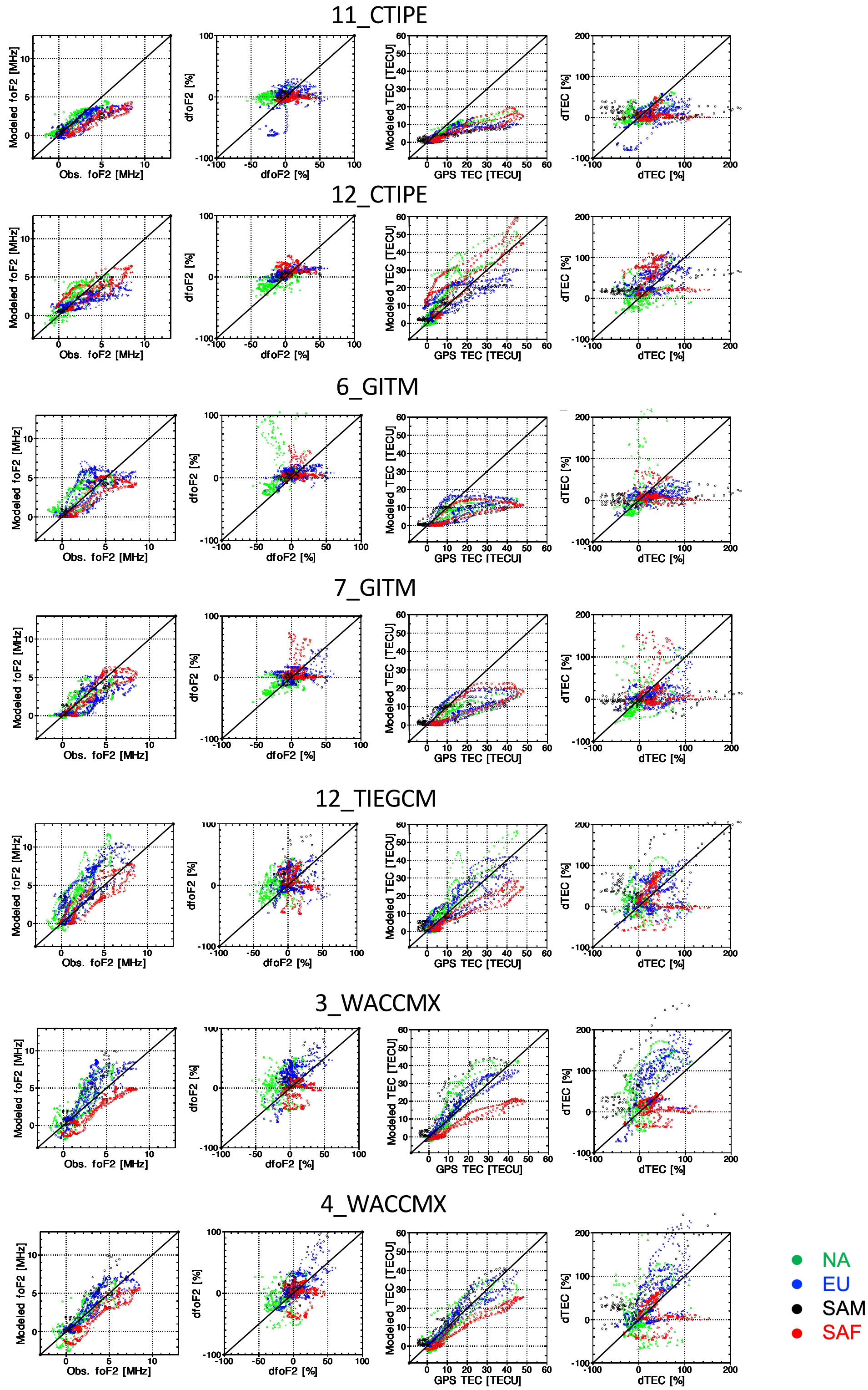
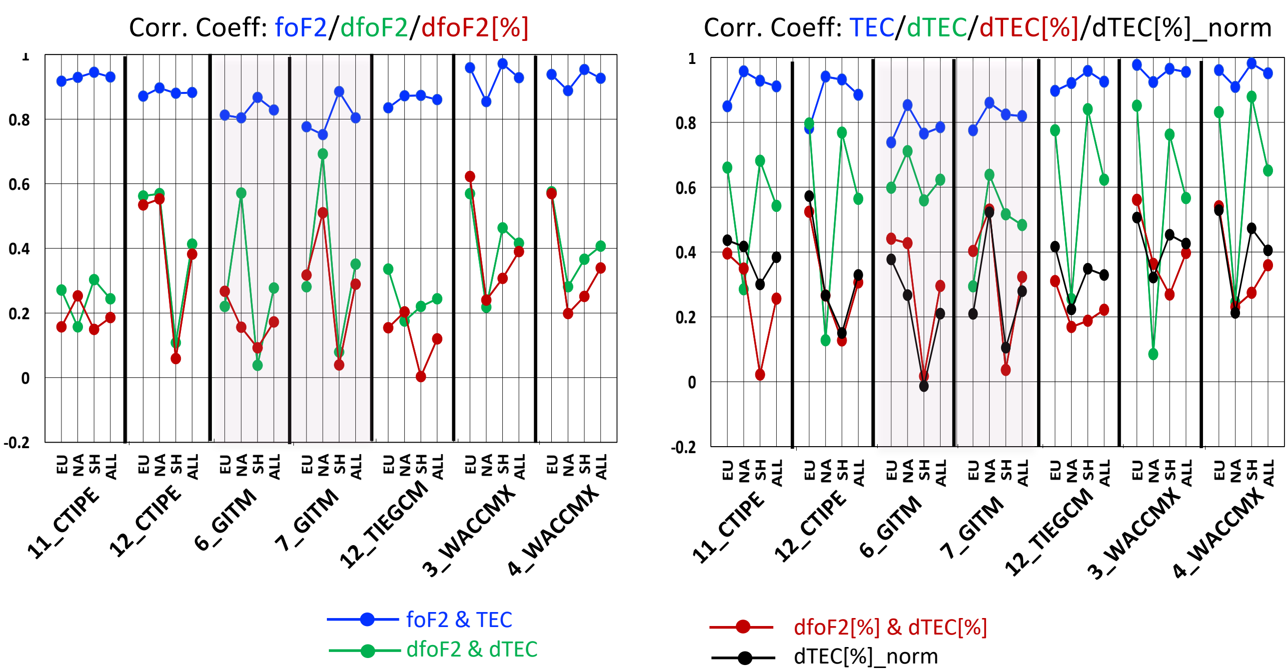
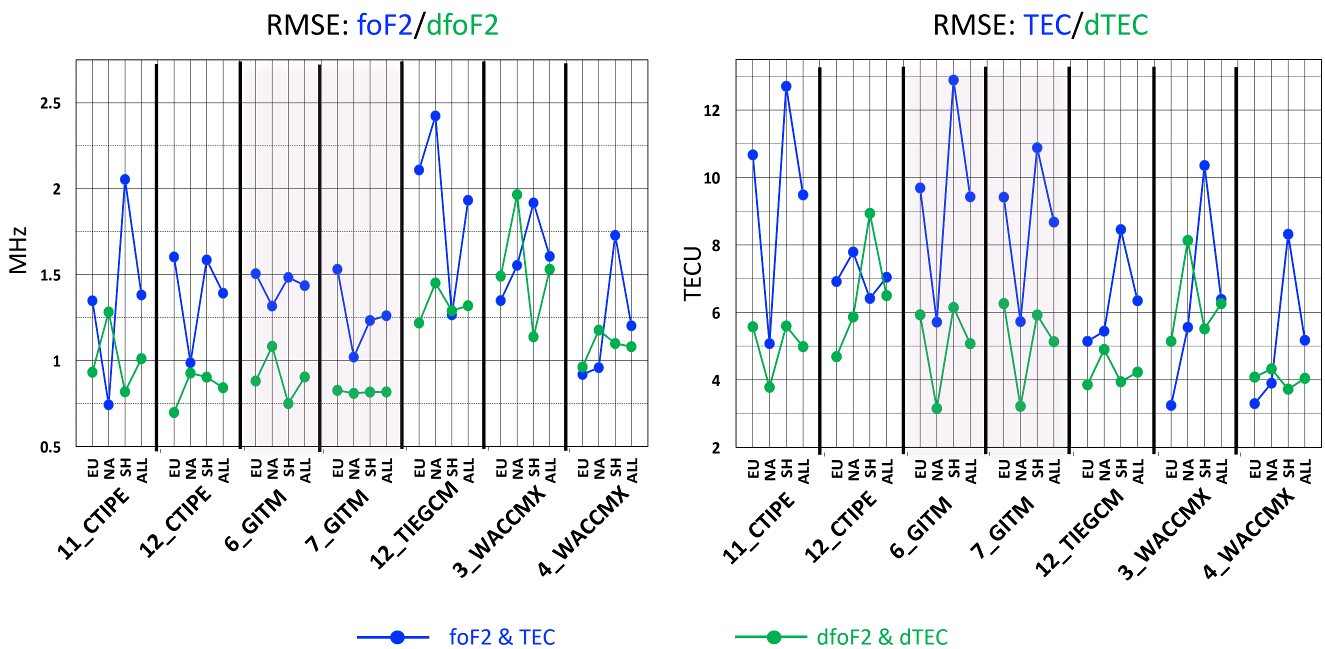

11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
Modeled foF2 [MHz]
Obs. foF2 [MHz]
dfoF2 [%]
Modeled TEC [TECU]
GPS TEC [TECU]
dTEC [%]
NA
EU
SAM
SAF

Figure 1



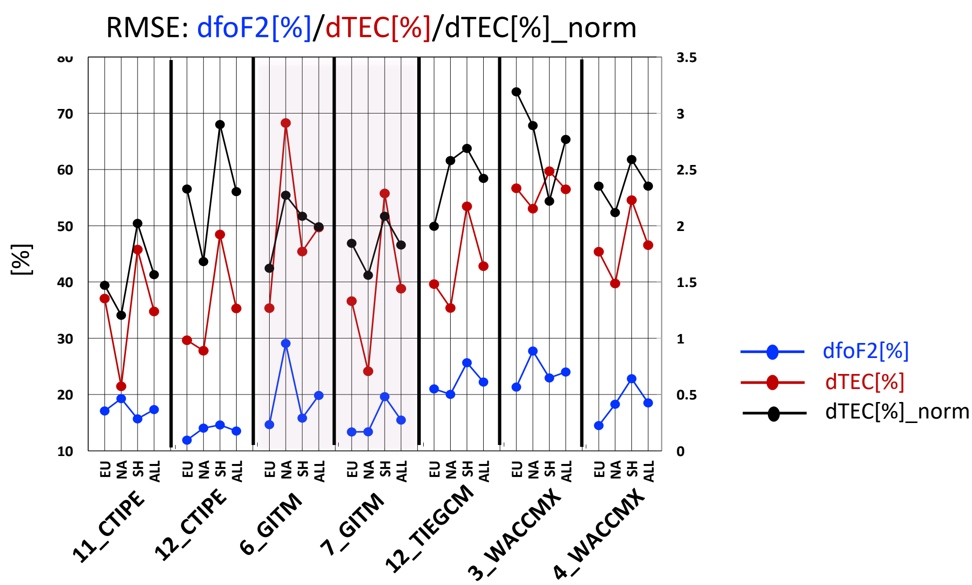
Corr. Coeff: foF2/dfoF2/dfoF2[%]
Corr. Coeff: TEC/dTEC/dTEC[%]/dTEC[%]_norm
1
0.8
0.6
0.4
0.2
0
-0.2
EU
NA
SH
ALL
11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
foF2 & TEC
dfoF2 & dTEC
dfoF2[%] & dTEC[%]
dTEC[%]_norm

Figure 2



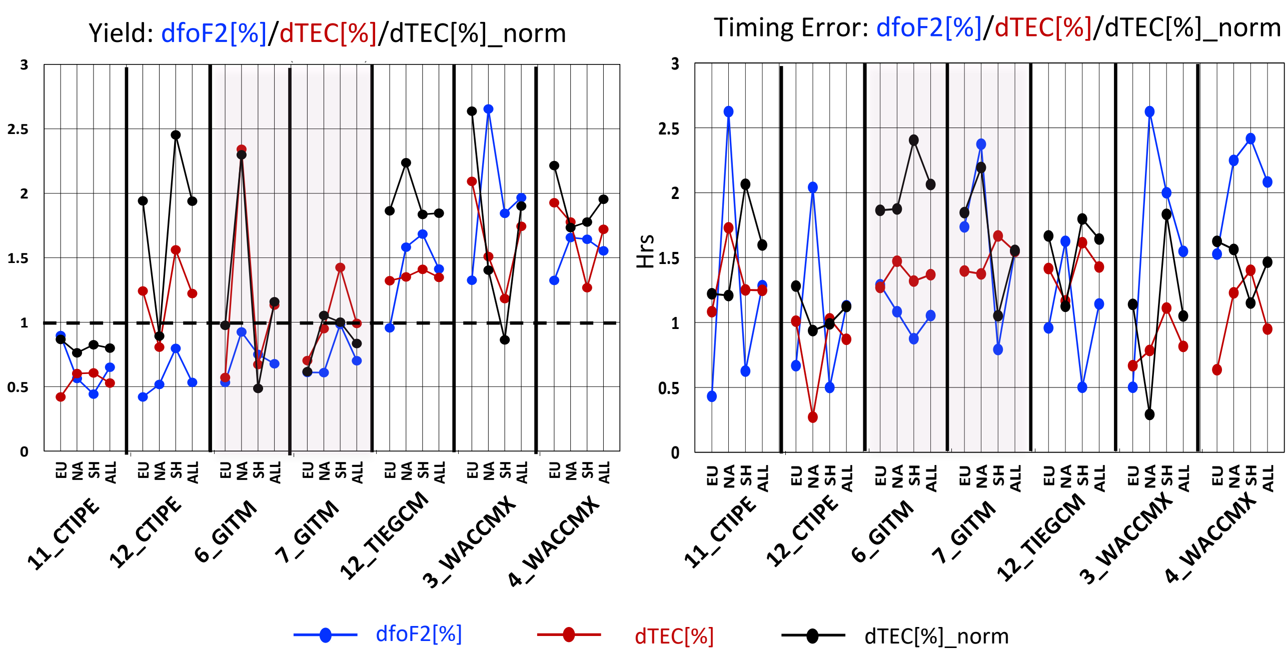
RMSE: foF2/dfoF2
MHz
2.5
2
1.5
1
0.5
RMSE: TEC/dTEC
TECU
12
10
8
6
4
2
EU
NA
SH
ALL
11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
foF2 & TEC
dfoF2 & dTEC

Figure 3



RMSE: dfoF2[%]/dTEC[%]/dTEC[%]_norm
[%]
80
70
60
50
40
30
20
10
3.5
3
2.5
2
1.5
1
0.5
0
EU
NA
SH
ALL
11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
dfoF2[%]
dTEC[%]
dTEC[%]_norm

Figure 4



Yield: dfoF2[%]/dTEC[%]/dTEC[%]_norm
Timing Error: dfoF2[%]/dTEC[%]/dTEC[%]_norm
Hrs
3
2.5
2
1.5
1
0.5
0
EU NA SH ALL
11_CTIPE
12_CTIPE
6_GITM
7_GITM
12_TIEGCM
3_WACCMX
4_WACCMX
dfoF2[%]
dTEC[%]
dTEC[%]_norm

Figure 5

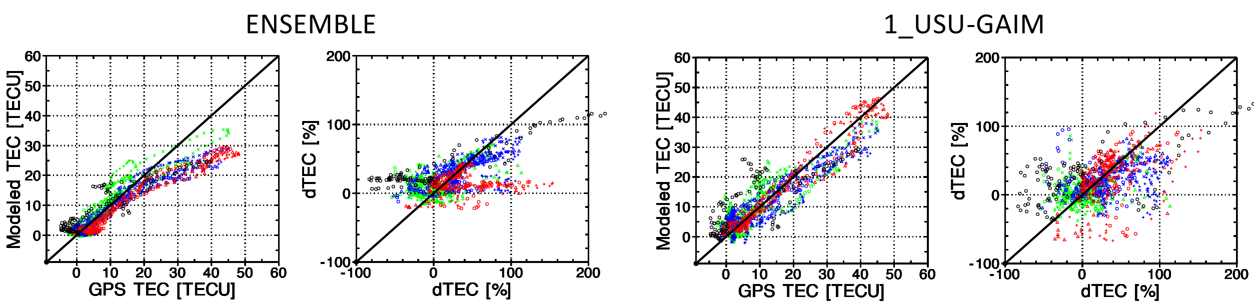


Figure 6

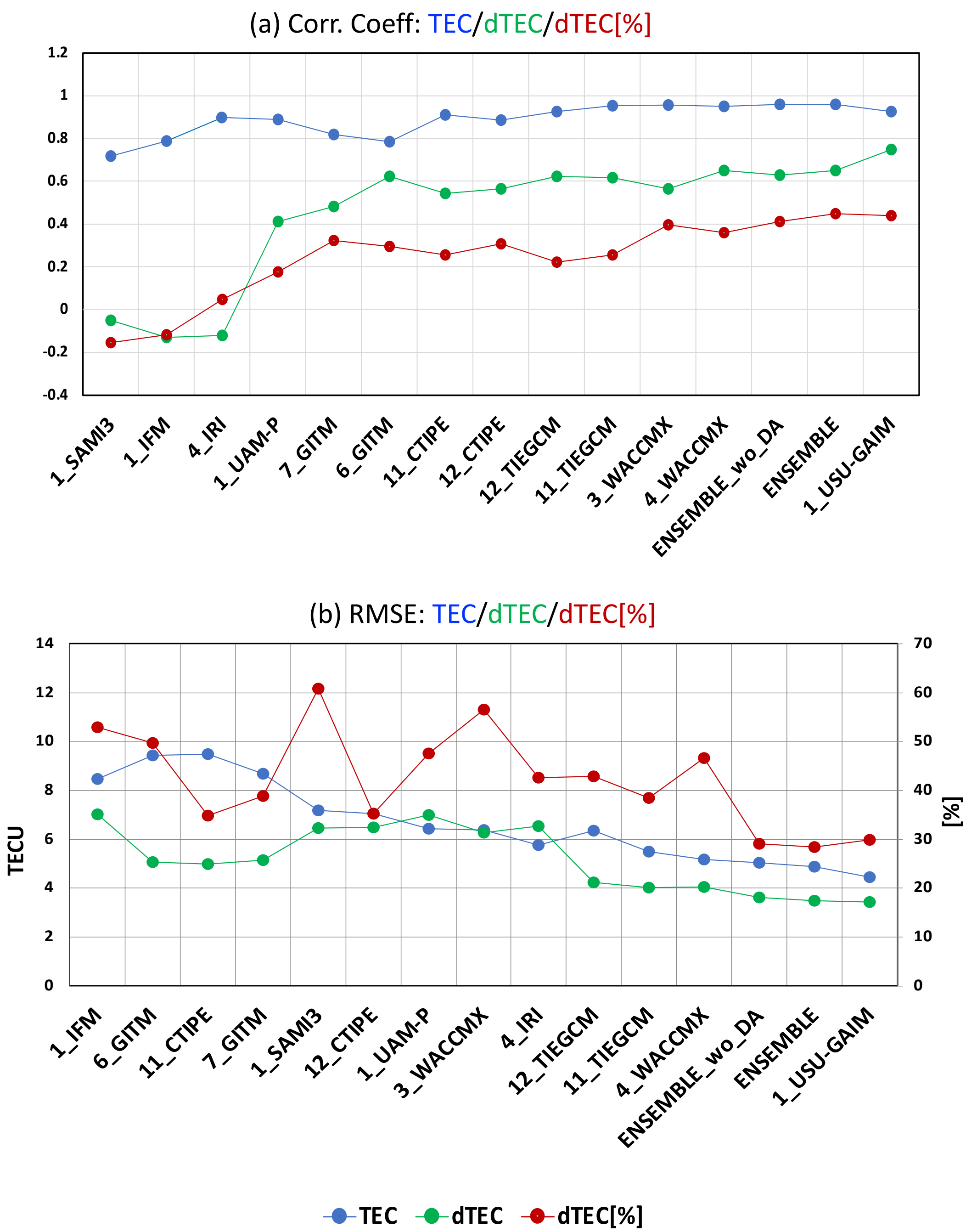
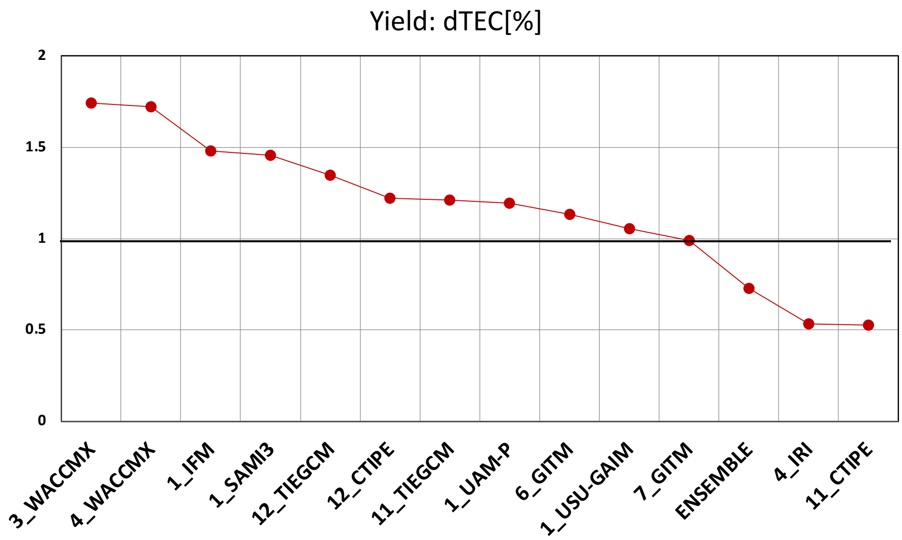
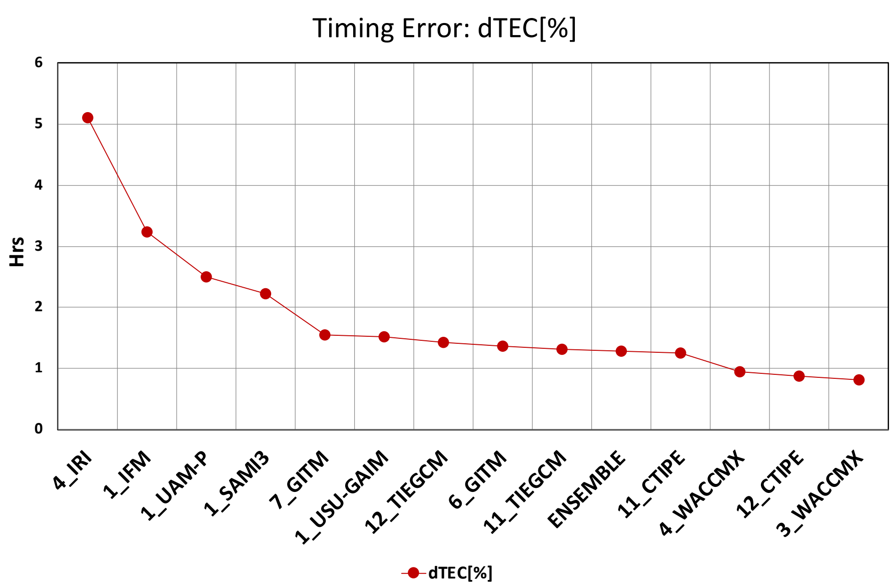

(a) Corr. Coeff: TEC/dTEC/dTEC[%]
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
1_SAMI3
1_IFM
4_IRI
1_UAM-P
7_GITM
6_GITM
11_CTIPE
12_CTIPE
12_TIEGCM
11_TIEGCM
3_WACCMX
4_WACCMX
ENSEMBLE_wo_DA
ENSEMBLE
1_USU-GAIM
(b) RMSE: TEC/dTEC/dTEC[%]
14
12
10
8
6
4
2
0
TECU
70
60
50
40
30
20
10
0
[%]
1_IFM
6_GITM
11_CTIPE
7_GITM
1_SAMI3
12_CTIPE
1_UAM-P
3_WACCMX
4_IRI
12_TIEGCM
11_TIEGCM
4_WACCMX
ENSEMBLE_wo_DA
ENSEMBLE
1_USU-GAIM
TEC
dTEC
dTEC[%]

Figure 7



Yield: dTEC[%]
2
1.5
1
0.5
0
3_WACCMX
4_WACCMX
1_IFM
1_SAMI3
12_TIEGCM
12_CTIPE
11_TIEGCM
1_UAM-P
6_GITM
1_USU-GAIM
7_GITM
ENSEMBLE
4_IRI
11_CTIPE



Timing Error: dTEC[%]
6
5
4
3
2
1
0
Hrs
4_IRI
1_IFM
1_UAM-P
1_SAMI3
7_GITM
1_USU-GAIM
12_TIEGCM
6_GITM
11_TIEGCM
ENSEMBLE
11_CTIPE
4_WACCMX
12_CTIPE
3_WACCMX
dTEC[%]

Figure 8

Figure 1. Scatter plots of the observed (*x* axis) and modeled (*y* axis) shifted foF2 and TEC (foF2* in the 1st , TEC* in the 3rd columns), and percentage change of foF2 and TEC (dfoF2[%] in the 2nd, dTEC[%] in the 4th columns) during the storm (03/17/2013) for all 12 locations grouped into North America (NA, green), Europe (EU, blue), South Africa (SAF, red), and South America (SAM, black)

Figure 2. Correlation Coefficients (CC) between modeled and observed foF2 (left panel) and TEC (right panel). Four CCs are displayed for each simulation: CC averaged over Europe (EU), North America (NA), Southern Hemisphere (SH refers to SAF and SAM combined), and all 12 locations, from left to right. Different colors denote different quantities. Blue denotes shifted foF2 and TEC, green and red the change and percentage changes, and black normalized percentage change. The closer the circles are to the horizontal line of 1, the better the model performances are.

Figure 3. Same as Figure 2 but for RMSE of shifted foF2 and TEC, and changes of foF2 and TEC

Figure 4. Same as Figure 2 but for RMSE of percentage change of foF2 and TEC, and normalized percentage change. Blue denotes dfoF2[%], red and black dTEC[%] and dTEC[%]_norm.

Figure 5. Same as Figure 2 but for Yield (ratio) and absolute of Timing Error (|TE| = |t_peak_model – t_peak_obs|)

Figure 6. Same as Figure 1 but for only TEC and dTEC[%] from the ensemble of the simulations (ENSEMBLE) and 1_USU-GAIM

Figure 7. Averaged CC (a) and RMSE (b) over all 12 locations of 13 simulations, the ensemble of them (ENSEMBLE), and the ensemble of 12 simulations excluding 1_USU-GAIM (ENSEMBLE_wo_DA). Blue denotes shifted TEC, green and red the change and percentage changes of TEC. CCs are plotted from the smallest to the largest (closer to 1) according to the average of the three averaged CC values of TEC, dTEC and dTEC[%]. RMSEs are plotted from the largest to the smallest according to the average RMSE for TEC and dTEC.

Figure 8. Yield and Timing Error of dTEC[%] for all 13 simulations and ENSEMBLE.

Table 1. Quantities and Skill Scores for Model-Data Comparison

Quantities and skill scores for model-data comparison	
Quiet time references	30-day median value at a given time: TEC_quiet(UT), 30 days consist of 15 days before (03/01-03/15/2013) and 15 days after (03/22-04/05/2013) the storm
Shifted TEC/foF2:	e.g., TEC*(doy, UT) = TEC(doy, UT) – minimum of TEC_quiet(UT)
TEC/foF2 changes w.r.t. the quiet time	e.g, dTEC(doy, UT)= TEC(doy, UT) –TEC_quiet (UT)
TEC/foF2 percentage changes w.r.t.the quiet time	e.g., dTEC[%](doy,UT) =100* dTEC(doy, UT)/TEC_quiet(UT)
Normalized Percentage changes of TEC	dTEC[%]_norm = (dTEC[%] -ave_dTEC[%])/std_dTEC[%]; ave_dTEC[%] is the average of dTEC[%] at a given time and at a given location over the quiet 30 days, std_dTEC[%] is the standard deviation of the average percentage change
Skill Scores	
CC	Correlation Coefficient
RMSE	Root-Mean-Square Error (=$\sqrt{\frac{\sum(x_{obs}-x_{mod})^2}{N}}$), where x_{obs} and x_{mod} are observed and modeled values
Yield	ratio of the peak of modeled percentage change to that of the observed one (=$\frac{(x_{mod})max}{(x_{obs})max}$)
Timing Error (TE)	difference between the modeled peak time and observed peak time: TE = t_peak_model – t_peak_obs

Table 2. Models used for this study

Model Setting ID	Model Version	Drivers			Upper boundary for TEC calculation/ Resolution
		Input data	Models used for thermosphere, tides from lower boundary, and high latitude electrodynamics		
Physics-based Coupled Ionosphere-Thermosphere Model					
			Tides	High Latitude Electrodynamics	
11_CTIPE[a]	CTIPe3.2 [*Codrescu et al.*, 2000; *Millward et al.*, 2001]	F10.7, ACE IMF data and solar wind speed and density, NOAA POES Hemispheric Power data	(2,2), (2,3), (2,4), (2,5), and (1,1) propagating tidal modes	Weimer-2005 high latitude electric potential [*Weimer*, 2005], Fuller-Rowell and Evans auroral precipitation [1987]	~2,000 km, 2° lat. × 18° long.
12_CTIPE[a]	CTIPe4.1		WAM [*Akmaev et al.,* 2011, *Fuller-Rowell et al.,* 2010] tides		
6_GITM[a]	GITM2.5 [*Ridley et al.*, 2006]	FISM solar EUV irradiance, ACE IMF data and solar wind speed and density	MSIS [*Hedin,* 1991] migrating diurnal and semidiurnal tides	Weimer-2005 high latitude electric potential, Ovation auroral precipitation [*Newell et al.,* 2009; 2011]	~600 km, 2.5° lat. × 5° long.
7_GITM	GITM21.11			Weimer-2005 high latitude electric potential, Fang's auroral precipitation [*Fang et al.,* 2013]	
12_TIE-GCM[a]	TIE-GCM2.0 [*Roble et al.*, 1988; *Richmond et al.*, 1992; *Solomon et al.*, 2012]	F10.7, Kp, OMNI IMF data and solar wind speed and density	GSWM [*Hagan et al.*, 1999] migrating diurnal and semidiurnal tides	Weimer-2005 high latitude electric potential, Roble and Ridley auroral precipitation [1987]	~600 km, 2.5° lat. × 2.5° long.
Whole Atmosphere Model					
3_WACCM-X	CESM2.2 [*Gettelman et al.,* 2019; *Liu et al.*, 2018]	F10.7, Kp, OMNI IMF data and solar wind speed and density	Heelis high latitude electric potential [*Heelis et al.,* 1982], Roble and Ridley auroral precipitation [1987]		~600 km, 1.9° lat. × 2.5° long.
4_WACCM-X			Weimer-2005 high latitude electric potential, Roble and Ridley auroral precipitation [1987]		

[a]The model results are submitted by the CCMC using the models hosted at the CCMC

Table 3. Number of locations where the models correctly predict negative or positive phase.

	Time Interval	11_CTIPE	12_CTIPE	6_GITM	7_GITM	12_TIE-GCM	3_WACCM-X	4_WACCM-X
dfoF2[%]	06–15UT	**8**	7	5	**9**	9	6	**10**
	15–22UT	**10**	6	7	**8**	7	7	**10**
dTEC[%]	06–15UT	9	**10**	**10**	**10**	7	**10**	9
	15–22UT	7	**10**	**12**	11	**10**	7	8

Table 4. Averaged RMSE over all 12 locations of the ensemble of newer versions (ENSEMBLE_new) of models (12_CTIPE, 7_GITM and 4_WACCM-X) driven by Weimer2005 electric potential model, the ensemble of older versions (ENSEMBLE_old) of models (11_CTIPE, 6_GITM and 12_TIE-GCM), and 1_USU-GAIM.

	TEC (TECU)	dTEC (TECU)	dTEC[%]
ENSEMBLE_old	6.6	4.1	33.4
ENSEMBLE_new	4.6	3.2	29.8
1_USU-GAIM	4.5	3.4	29.9

AGU PUBLICATIONS

Space Weather: The International Journal of Research and Applications

Supporting Information for

Validation of Ionospheric Specifications During Geomagnetic Storms: TEC and foF2 during the 2013 March Storm Event-II

J. S. Shim[1], I.-S. Song[1], G. Jee[2], Y.-S. Kwak[3], I. Tsagouri[4], L. Goncharenko[5], J. McInerney[6], A. Vitt[6], L. Rastaetter[7], J. Yue[7,8], M. Chou[7,8], M. Codrescu[9], A. J. Coster[5], M. Fedrizzi[9], T. J. Fuller-Rowell[9], A. J. Ridley[10], S. C. Solomon[6]

[1]Department of Atmospheric Sciences, Yonsei University, Seoul, South Korea,
[2]Division of Atmospheric Sciences, Korea Polar Research Institute, Incheon, South Korea
[3]Space Science Division, Korea Astronomy and Space Science Institute, Daejeon, South Korea
[4]National Observatory of Athens, Penteli, Greece,
[5]Haystack Observatory, Westford, MA, USA,
[6]High Altitude Observatory, NCAR, Boulder, CO, USA,
[7]NASA GSFC, Greenbelt, MD, USA,
[8]Catholic University of America, Washington, DC, USA,
[9]NOAA SWPC, Boulder, CO, USA,
[10]Space Physics Research Laboratory, Univ. of Michigan, Ann Arbor, MI, USA

Contents of this file

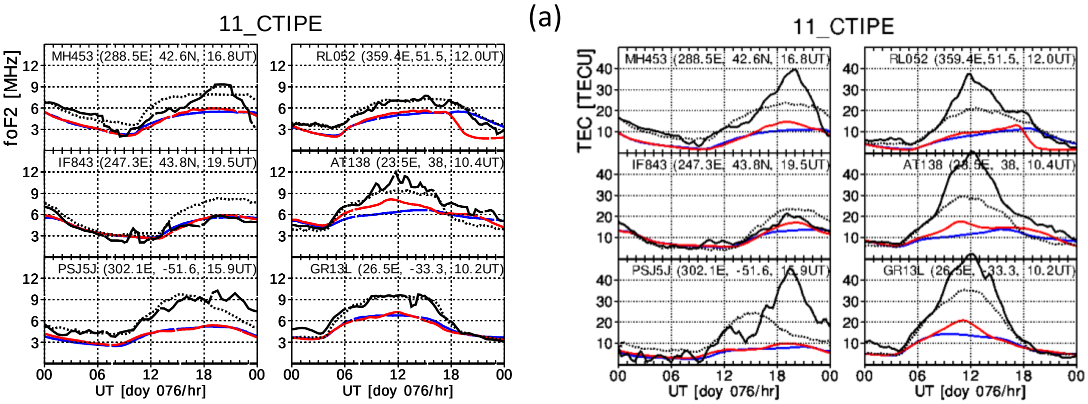
Figures S1

Introduction

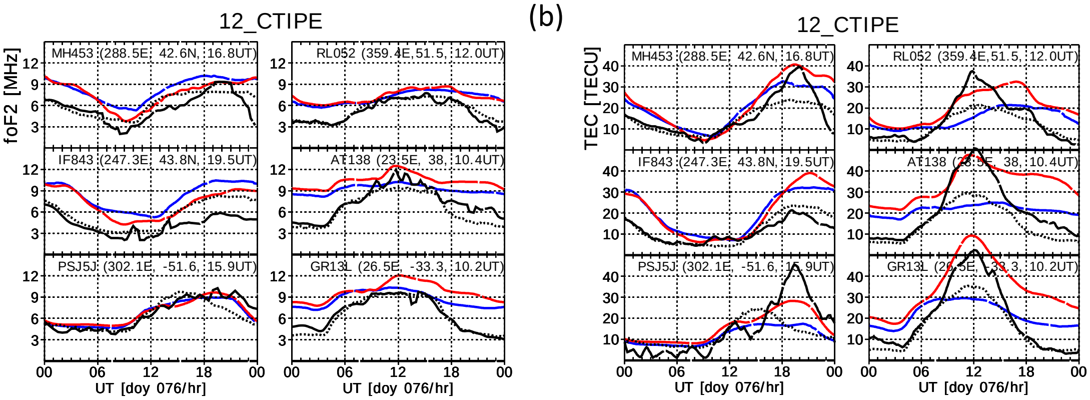
This supporting information file includes:

1. Figure S1: Comparison foF2 and TEC data with modeled values

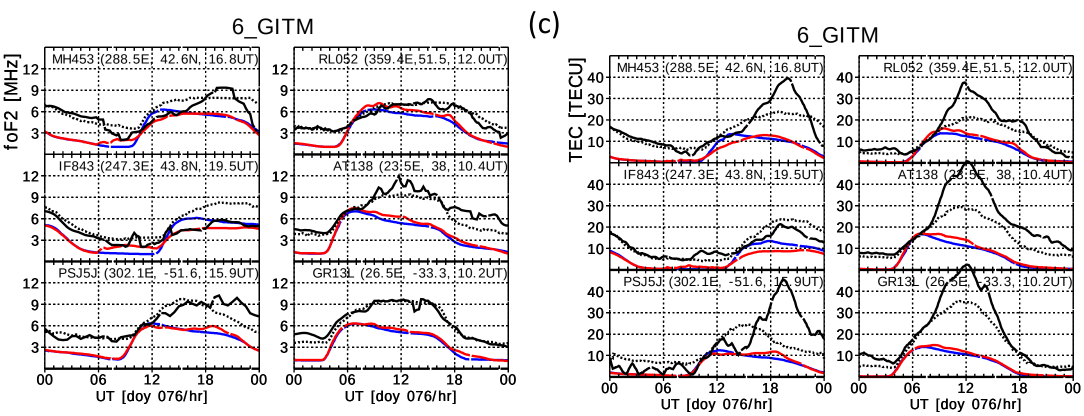
(a)


11_CTIPE
foF2 [MHz]
TEC [TECU]
MH453 (288.5E, 42.6N, 16.8UT)
RL052 (359.4E,51.5, 12.0UT)
IF843 (247.3E, 43.8N, 19.5UT)
AT138 (23.5E, 38, 10.4UT)
PSJ5J (302.1E, -51.6, 15.9UT)
GR13L (26.5E, -33.3, 10.2UT)
UT [doy 076/hr]

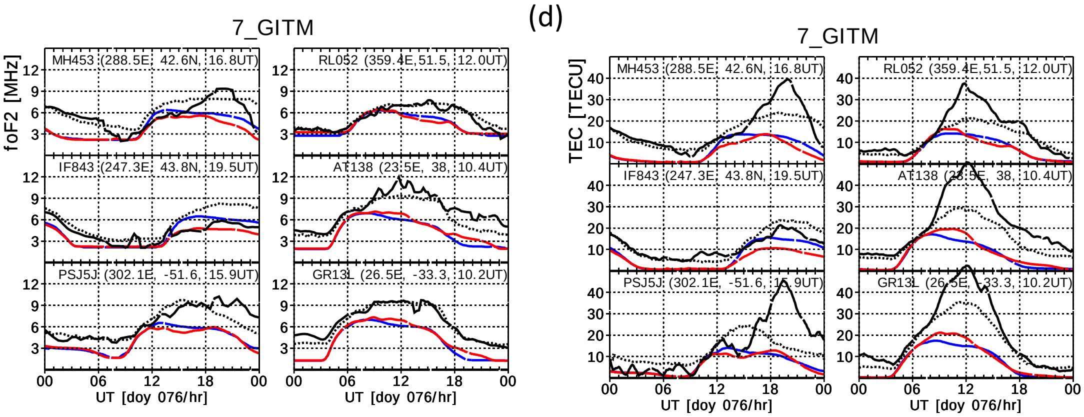
(b)


12_CTIPE
foF2 [MHz]
TEC [TECU]
MH453 (288.5E, 42.6N, 16.8UT)
RL052 (359.4E,51.5, 12.0UT)
IF843 (247.3E, 43.8N, 19.5UT)
AT138 (23.5E, 38, 10.4UT)
PSJ5J (302.1E, -51.6, 15.9UT)
GR13L (26.5E, -33.3, 10.2UT)
UT [doy 076/hr]

(c)


6_GITM
foF2 [MHz]
TEC [TECU]
MH453 (288.5E, 42.6N, 16.8UT)
RL052 (359.4E,51.5, 12.0UT)
IF843 (247.3E, 43.8N, 19.5UT)
AT138 (23.5E, 38, 10.4UT)
PSJ5J (302.1E, -51.6, 15.9UT)
GR13L (26.5E, -33.3, 10.2UT)
UT [doy 076/hr]

(d)


7_GITM
foF2 [MHz]
TEC [TECU]
MH453 (288.5E, 42.6N, 16.8UT)
RL052 (359.4E,51.5, 12.0UT)
IF843 (247.3E, 43.8N, 19.5UT)
AT138 (23.5E, 38, 10.4UT)
PSJ5J (302.1E, -51.6, 15.9UT)
GR13L (26.5E, -33.3, 10.2UT)
UT [doy 076/hr]

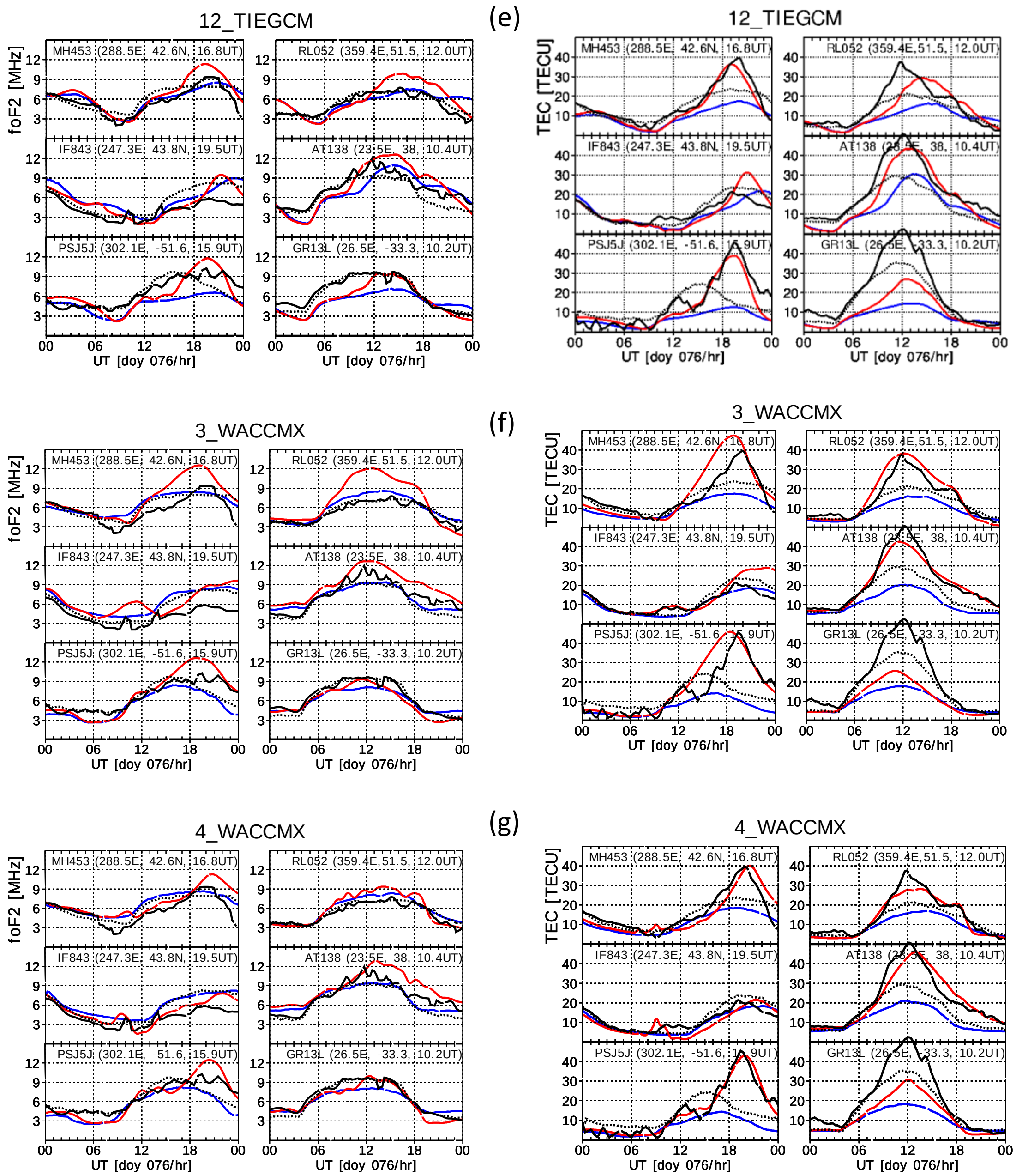


Figure S1. Comparison foF2 and TEC data with modeled values: (a) 11_CTIPE, (b) 12_CTIPE, (c) 6_GITM, (d) 7_GITM, (e) 12_TIEGCM, (f) 3_WACCM-X, and (g) 4_WACCM-X. In each plot, foF2 in the first two columns and TEC in the other two. Black solid and dotted lines denote observed storm time values and quiet-time reference (30-day median) respectively. Red and blue curves denote modeled storm time values and 30-day median.