

RESEARCH ARTICLE

10.1002/2017JA025109

Key Points:

- Synthetic FORMOSAT-7/COSMIC-2 RO sTEC are assimilated into a coupled model of the thermosphere, ionosphere, and plasmasphere by the EnSRF
- Observing system simulation experiments are used to examine the effects of key EnSRF parameters on the ionospheric specification in detail
- The ensemble size larger than 70 and the horizontal covariance localization with a length scale of 5,000 km is required to improve the result

Supporting Information:

- Supporting Information S1

Correspondence to:

J.-Y. Liu,
jyliu@jupiter.ss.ncu.edu.tw

Citation:

Hsu, C.-T., Matsuo, T., Yue, X., Fang, T.-W., Fuller-Rowell, T., Ide, K., & Liu, J.-Y. (2018). Assessment of the impact of FORMOSAT-7/COSMIC-2 GNSS RO observations on midlatitude and low-latitude ionosphere specification: Observing system simulation experiments using Ensemble Square Root Filter. *Journal of Geophysical Research: Space Physics*, 123, 2296–2314. <https://doi.org/10.1002/2017JA025109>







Received 11 DEC 2017

Accepted 21 FEB 2018

Accepted article online 1 MAR 2018

Published online 23 MAR 2018

Assessment of the Impact of FORMOSAT-7/COSMIC-2 GNSS RO Observations on Midlatitude and Low-Latitude Ionosphere Specification: Observing System Simulation Experiments Using Ensemble Square Root Filter

C.-T. Hsu^{1,2}, T. Matsuo^{2,3} , X. Yue⁴ , T.-W. Fang^{2,5} , T. Fuller-Rowell⁴ , K. Ide⁶ , and J.-Y. Liu¹ 

¹Institute of Space Science, National Central University, Taoyuan, Taiwan, ²Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA, ³Aerospace Engineering Sciences, University of Colorado, Boulder, CO, USA, ⁴Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China, ⁵Satellite Research Center, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, ⁶Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA

Abstract The Formosa Satellite-7/Constellation Observing System for Meteorology, Ionosphere, and Climate-2 (FORMOSAT-7/COSMIC-2) Global Navigation Satellite System radio occultation (RO) payload can provide global observations of slant total electron content (sTEC) with an unprecedentedly high spatial temporal resolution. Recently, a new ionospheric data assimilation system, the Community Gridpoint Statistical Interpolation (GSI) Ionosphere, is constructed with the National Oceanic and Atmospheric Administration GSI Ensemble Square Root Filter and the Global Ionosphere Plasmasphere and the Thermosphere Ionosphere Electrodynamic General Circulation Model. The paper demonstrates the capability of the GSI Ionosphere to improve the ionospheric specification and make a quantitative assessment of the impact of FORMOSAT-7/COSMIC-2 RO data on the ionospheric observing system simulation experiments conducted to calibrate key Ensemble Square Root Filter parameters that control detrimental effects of the sampling errors, particularly on the ensemble-based estimation of the correlation between observations and model states, in order to yield high-quality assimilation analysis. Results from the observing system simulation experiments show that (1) an ensemble size larger than 70 is recommended for assimilation of RO sTEC data with the GSI Ionosphere and (2) localizing the impact of observations around the tangent points in the horizontal direction with a length scale of 5,000 km is effective in improving assimilation analysis quality. Assimilation of sTEC data from FORMOSAT-7/COSMIC-2 can considerably improve the global ionospheric specification through the application of the GSI Ionosphere. The GSI Ionosphere can provide instantaneous global pictures of the ionosphere variability and help characterize day-to-day variability of the ionosphere and deepen our understanding of the observed day-to-day variability.

1. Introduction

As exemplified by the success of numerical weather prediction, data assimilation has attracted great attention as a promising approach to integrating geospace observing capabilities with a numerical model of the ionosphere to improve the specification and forecasting of ionospheric weather. Data assimilation is a powerful technique that can optimally combine observations with a numerical model to help initialize model states and estimate inadequately specified model parameters.

Total electron content (TEC) is one of the most valued data types for ionospheric data assimilation. From the phase and pseudo range measurements of the Global Navigation Satellite System (GNSS) signals received by ground-based and satellite-based GNSS receivers, the electron density integrated along the radio path between GNSS satellites and receivers through the ionosphere and plasmasphere can be calculated and is referred to as slant TEC (sTEC). Comparing with ground-based TEC, satellite-based TEC data are evenly distributed over ocean and land areas. A radio occultation (RO) event occurs when the Earth is located between GNSS and low Earth orbit satellites, while the GNSS raypath between GNSS transmitters and receivers passes through the ionosphere and plasmasphere. Otherwise, the antenna at low Earth orbit receives the

signal that travels only through the plasmasphere. In this study, we will focus on RO sTEC observations, in particular, expected to be obtained from the Formosa Satellite Mission 7/Constellation Observing System for Meteorology, Ionosphere and Climate 2 (FORMOSAT-7/COSMIC-2) mission. Thanks to the success of the FORMOSAT-3/COSMIC mission, the follow-on FORMOSAT-7/COSMIC-2 mission is originally consist of six low-inclination-angle (24° – 28.5°) orbit satellites and six high-inclination-angle orbit (72°) satellites. The six low-inclination-angle orbit satellites will be launched in 2018. Although the mission of six high-inclination-angle orbit satellites is cancelled, it is expected that there will very likely an alternative source available from high latitude constellation from commercial providers. The main payload, TriG GNSS RO System, is capable of profiling the ionosphere with a great accuracy. The FORMOSAT-7/COSMIC-2 low-inclination angle satellites are expected to provide a dense spatial and temporal coverage of high-quality sTEC observations evenly distributed on low latitude and midlatitude (Yue, Schreiner, Kuo, et al., 2014; Yue, Schreiner, Pedatella, et al., 2014).

In the past decade, considerable efforts have been made to assimilate TEC data into numerical ionospheric models (e.g., Chartier et al., 2016; Chen et al., 2016; Scherliess et al., 2006; Schunk et al., 2004; Wang et al., 2004). One of the most widely recognized ionospheric data assimilation systems is the Global Assimilative Ionospheric Model (GAIM) developed by a joint effort of University of Southern California and Jet Propulsion Laboratory (USC-JPL GAIM). The USC-JPL GAIM can assimilate multiple types of data into a numerical ionospheric model. The model covers the altitude range from 100 to 1,500 km. The plasma density in this model is calculated along the geomagnetic field lines, and the thermospheric states and electric fields along with other drivers such as solar extreme ultraviolet (EUV), on the other hand, are parameterized or specified by empirical models. In the USC-JPL GAIM, two data assimilation schemes are used to update different variables. The four-dimensional variation data assimilation method is used to estimate model drivers, and a band-limited Kalman Filter is used to estimate the plasma density model state (Hajj et al., 2004). Hajj et al. (2004) have successfully assimilated ground-based TEC data into the USC-JPL GAIM by using this band-limited Kalman Filter. Komjathy et al. (2010) have assimilated both ground-based Global Positioning System (GPS) TEC and the FORMOSAT-3/COSMIC TEC data into the USC-JPL GAIM. Their results show assimilating satellite-based TEC data helps improve vertical electron density specifications.

The Global Assimilation of Ionospheric Measurement (GAIM) developed by Utah State University (USU GAIM) is another well-recognized ionospheric data assimilation system. This system has two different data assimilation approaches, the USU GAIM-GM (Scherliess et al., 2006; Schunk et al., 2004, 2016) and the USU GAIM-FP (Scherliess, Thompson, & Schunk, 2009; Schunk et al., 2016). The model used in the USU GAIM-GM is the Ionosphere Forecast Model (IFM), and the data assimilation scheme is Gauss-Markov Kalman filter. The IFM covers the altitude range from 90 to 1,400 km, and the plasma density is calculated along the geomagnetic field lines. In IFM, thermospheric compositions, temperatures, and winds as well as electric field and precipitation patterns are specified by empirical models (Scherliess et al., 2006; Schunk et al., 2004). On the other hand, the model used in the USU GAIM-FP is the Ionosphere-Plasmasphere Model (IPM) that covers the altitude range from 90 to 30,000 km. As for the IFM, the IPM needs thermospheric state variables and other external drivers to be specified by empirical models but can be adjusted by the data assimilation system. In addition, the electron and ion temperature from empirical models are necessary because energy equations are not solved in the IPM. The data assimilation scheme used in the USU GAIM-FP is ensemble Kalman filter (Scherliess, Thompson, & Schunk, 2009; Schunk et al., 2004). Both the USU GAIM-GM and the USU GAIM-FP can assimilate TEC data into models adequately (Scherliess et al., 2006; Scherliess, Thompson, & Schunk, 2009; Schunk et al., 2016).

Recently, another data assimilation system has been built for the National Center for Atmospheric Research (NCAR) Thermosphere Ionosphere Electrodynamics General Circulation Model (TIE-GCM) (Richmond, Ridley, & Roble, 1992) using the Ensemble Square Root Filter (EnSRF) implemented in Data Assimilation Research Testbed (DART) (Anderson, 2001, 2003). The TIE-GCM is a three-dimensional, physical-based model that can self-consistently simulate the coupled processes of ionosphere and thermosphere in hydrostatic pressure coordinates. The DART is a flexible ensemble data assimilation software framework with various options for filtering methods. Unlike the GAIM systems, the DART/TIE-GCM has been designed to take a full advantage of thermosphere-ionosphere coupling in both analysis and forecast steps of EnSRF (e.g., Hsu et al., 2014; Lee et al., 2012; Matsuo & Araujo-Pradere, 2011; Matsuo, Lee, & Anderson, 2013). The DART/TIE-GCM has also been used to assimilate ground-based TEC data successfully during ionospheric storm periods (e.g., Chartier et al., 2016; Chen et al., 2016). Because the upper boundary of the TIE-GCM is too low to represent contributions of

the topside ionosphere and plasmasphere electron density to sTEC adequately, Chen et al. (2016) and Chartier et al. (2016) had to extrapolate electron densities above the model upper boundary to assimilate vertical TEC into the DART/TIE-GCM. Vertical TEC is the vertical integration of the electron density in the direction perpendicular to the ground and can be geometrically converted from sTEC under some assumptions. To avoid the errors introduced by vertical extrapolation of the model electron densities, it is desirable to use a model that includes the ionosphere and plasmasphere for assimilation of sTEC data.

In order to take an advantage of ionosphere-thermosphere coupling in data assimilation methods and to overcome limitations of earlier works by the DART/TIE-GCM, in this study, we present a new approach for assimilation of satellite-based RO sTEC data. The study focuses on assessment of the impact of sTEC observations from the upcoming FORMOSAT-7/COSMIC-2 low-inclination satellite constellation on the low-latitude and midlatitude ionospheric specification.

2. Data Assimilation System

In this study, a coupled model of the thermosphere, ionosphere, and plasmasphere developed as result of collaboration between NCAR and National Oceanic and Atmospheric Administration (NOAA) has been incorporated in an ensemble-based data assimilation scheme, which is part of the Gridpoint Statistical Interpolation (GSI) data assimilation system operationally used for numerical weather prediction at NOAA in order to build a new ionospheric data assimilation system. This system is hereafter referred to as the GSI Ionosphere.

2.1. Model

The model used in this study is a fully coupled model of the Global Ionosphere Plasmasphere (GIP) and the TIE-GCM (Pedatella et al., 2011). While the GIP simulates the ionosphere and plasmasphere processes, the TIE-GCM solves for the thermospheric processes including the electrodynamic processes. In the following, we refer to this coupled model as the GIP/TIE-GCM.

The GIP is developed from the ionosphere and plasmasphere part of the Coupled Thermosphere-Ionosphere-Plasmasphere Model (Millward et al., 2001). The GIP solves the continuity, momentum, and energy equations for plasma along geomagnetic field prescribed according to the International Geomagnetic Reference Field using the apex coordinate system (Richmond, 1995). In the GIP, the distribution of atomic oxygen and hydrogen ion densities are determined with consideration of the transport and diffusion processes. Other ion species solved by using atomic oxygen ion densities from the flux-tube solver and assume chemical equilibrium, a balance between production and loss. Therefore, both atomic oxygen and hydrogen ions are primary prognostic model state variables that are dynamically evolved from the previous model time step to the next in the GIP. The GIP model consists of two model domains: the low-latitude and midlatitude regions and the high-latitude region. The GIP fluxtubes at a given magnetic longitude are distributed with respect to L shell. The boundary between low-latitude and midlatitude and high-latitude regions is fixed at $L = 4$. For the low-latitude and midlatitude regions, the GIP solves the plasma along closed fluxtubes that move perpendicular to the magnetic field (B) in the magnetic meridional/vertical direction by $E \times B$ and parallel to B by ambipolar diffusion (Millward et al., 2001). The altitude range of low-latitude and midlatitude parts of the GIP is approximately from 90 to 19,000 km, which covers the most of GNSS raypath of FORMOSAT-7/COSMIC-2 RO that traverses through the ionosphere and plasmasphere. On the other hand, the open fluxtubes in the high-latitude region are cut off at around 10,000 km in altitude, and therefore, the altitude range of the GIP high-latitude region is approximately from 100 to 10,000 km.

As mentioned above, the TIE-GCM solves for the thermospheric states, including electrodynamic, in the fixed pressure coordinates (Richmond, Ridley, & Roble, 1992). The horizontal resolution of the TIE-GCM version used for this study is $5^\circ \times 5^\circ$, and the vertical resolution is two levels per scale height. The altitude of the lower boundary of the TIE-GCM is approximately 97 km, and the upper boundary ranges from 400 to 700 km depending on solar activity levels. By using the GIP/TIE-GCM, we are able to account for ionosphere-thermosphere coupling in the process of data assimilation. The main drivers of the GIP/TIE-GCM include F10.7 index (F107), cross-tail potential drop (CP), auroral hemispheric power (HP), and atmospheric tides. F107 represents the solar EUV level that determines the photoionization rates, photo-dissociation rates, and heating rates of the neutral and ionized species in the model. HP and CP

indexes represent the magnitude of auroral particle precipitation and the ionospheric convective electric fields imposed from the magnetosphere. The atmospheric tides control the lower boundary conditions of the model.

2.2. Data Assimilation Methods

The data assimilation system under consideration is composed of an analysis step and a forecast step. In the analysis step, a selected set of the model state variables are updated through assimilation of observations. In the forecast step, updated state variables are fed back to the model and used as initial conditions to forecast the future state. In a data assimilation system, cycling of these two steps is carried out over an extended period. This is the so-called data assimilation cycle.

In this study, the atomic oxygen ion density and electron density on the model grid are selected to be estimated and updated during the analysis step. The electron density is an observed variable, but in the GIP/TIE-GCM, the electron density is recomputed as a sum of the atomic and molecular ion species at each model time step. In fact, the atomic oxygen ion is one of the main prognostic model state variables and the dominant ion species in the F region, whose number density is largely equal to the electron number density. The data assimilation scheme used in the analysis step in this study is the EnSRF developed by Whitaker and Hamill (2002) implemented in NOAA's GSI data assimilation system.

The EnSRF can be presented as a modification to the traditional Kalman filter (Kalman, 1960) and to the ensemble Kalman filter (Evensen, 1994). Following the standard notation used in the atmospheric data assimilation (e.g., Ide et al., 1997), let \mathbf{x}^a and \mathbf{x}^b be an m -dimensional vector of the updated state variables and forecast state variables, respectively, and \mathbf{y}^o be a p -dimensional vector of observational variables. \mathbf{P}^a and \mathbf{P}^b here represent an $m \times m$ analysis error covariance matrix and forecast error covariance matrix, respectively, and \mathbf{R} denotes a $p \times p$ observational error covariance matrix. The Kalman gain matrix and the gain used to update deviations are denoted as \mathbf{K} and $\tilde{\mathbf{K}}$, respectively. The forward operator that converts the state in the model space to the observation space is represented by a $p \times m$ matrix, \mathbf{H} . Following the formulation presented in Whitaker and Hamill (2002), a prime here denotes the deviation from the ensemble mean and an overbar denotes the ensemble mean. In all ensemble-based Kalman filters, including EnSRF, sample estimates of \mathbf{P}^b do not need to be explicitly computed and stored. Instead, the terms $\mathbf{P}^b \mathbf{H}^T$ and $\mathbf{H} \mathbf{P}^b \mathbf{H}^T$ are computed from model ensemble as shown below. Under the assumption that observational errors are not correlated, observations can be assimilated sequentially one by one. In this serial application of the analysis update, \mathbf{K} and $\tilde{\mathbf{K}}$ become vectors, and $\mathbf{H} \mathbf{P}^b \mathbf{H}^T$ and \mathbf{R} become scalars. This makes the filter implementation computationally more efficient. For a set of N model ensembles and one observation ($p = 1$), the state and covariance update equation in EnSRF are given as

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \mathbf{K}(\mathbf{y}^o - \mathbf{H}\bar{\mathbf{x}}^b) \quad (1)$$

$$\mathbf{x}'_n{}^a = \mathbf{x}'_n{}^b + \tilde{\mathbf{K}}(-\mathbf{H}\mathbf{x}'_n{}^b) \quad (2)$$

$$\mathbf{K} = [\boldsymbol{\rho}^b \circ (\mathbf{P}^b \mathbf{H}^T)] [\boldsymbol{\rho}^o \circ (\mathbf{H} \mathbf{P}^b \mathbf{H}^T) + \mathbf{R}]^{-1} \quad (3)$$

$$\tilde{\mathbf{K}} = \alpha \mathbf{K} \quad (4)$$

$$\alpha = \left(1 + \sqrt{\frac{\mathbf{R}}{\mathbf{H} \mathbf{P}^b \mathbf{H}^T + \mathbf{R}}} \right)^{-1} \quad (5)$$

$$\mathbf{P}^b \mathbf{H}^T \sim \frac{1}{(N-1)} \sum_{n=1}^N (\mathbf{x}_n^b - \bar{\mathbf{x}}^b) [\mathbf{H}(\mathbf{x}_n^b - \bar{\mathbf{x}}^b)]^T \quad (6)$$

$$\mathbf{H} \mathbf{P}^b \mathbf{H}^T \sim \frac{1}{(N-1)} \sum_{n=1}^N [\mathbf{H}(\mathbf{x}_n^b - \bar{\mathbf{x}}^b)] [\mathbf{H}(\mathbf{x}_n^b - \bar{\mathbf{x}}^b)]^T. \quad (7)$$

where n is an index for ensemble member ($n = 1, \dots, N$), $\boldsymbol{\rho}^b$ and $\boldsymbol{\rho}^o$ are $m \times p$ and $p \times p$ matrices of the covariance localization function, ρ (explain later), and \circ denotes the element-wise multiplication (i.e., Schur product). The ensemble means of the updated and forecast state variables are given as $\bar{\mathbf{x}}^a = \sum_{n=1}^N \mathbf{x}_n^a$ and $\bar{\mathbf{x}}^b = \sum_{n=1}^N \mathbf{x}_n^b$. The deviations of each updated and forecast state variables from the ensemble mean are given as $\mathbf{x}'_n{}^a = \mathbf{x}_n^a - \bar{\mathbf{x}}^a$ and $\mathbf{x}'_n{}^b = \mathbf{x}_n^b - \bar{\mathbf{x}}^b$ for each ensemble member. Please note that since the observation under

consideration in this study is sTEC, \mathbf{H} represents an operation that computes sTEC values from electron density values on the model grid and will be discussed in detail in the next subsection. In this kind of filter implementation, the sampling errors, which originate from the use of a finite size of the model ensemble ($N \ll m$) that detrimentally impacts on the estimation of $\mathbf{P}^b \mathbf{H}^T$ and $\mathbf{H} \mathbf{P}^b \mathbf{H}^T$, need to be addressed somehow. As shown later, this detrimental impact can be mitigated by the covariance localization. In the following section, the *prior* refers to a probability distribution of the model forecast ensemble before being updated by data assimilation and the *posterior* refers to a probability distribution of the model ensemble after the update in analysis step.

Commonly used auxiliary methods for adjusting \mathbf{P}^b to correct the issues associated with sampling errors include covariance inflation (e.g., Anderson & Anderson, 1999) and covariance localization (e.g., Hamill et al., 2001; Houtekamer & Mitchell, 2001). An underdispersed model ensemble leads to an insufficient variance in \mathbf{P}^a , which in turn causes filter divergence. Covariance inflation artificially inflates the sample variance of the model ensemble by effectively pushing an ensemble member away from the ensemble mean. The GSI-EnSRF uses the relation to prior spread (Whitaker & Hamill, 2012) that inflates the posterior variance by multiplying an inflation factor, γ , to each model ensembles perturbation.

$$\gamma = w \left(\frac{\sigma_b - \sigma_a}{\sigma_a} \right) + 1 \quad (8)$$

where σ_a is the posterior standard deviation, σ_b is the prior standard deviation, and w is the weighting factor for inflation. If $w = 1$, the posterior variance is the same as prior variance. If $w = 0$, there is no inflation.

On the other hand, the correlation estimated from a small number of the model ensemble often leads to spurious correlation, especially at large-lag distance. To suppress this spurious correlation in \mathbf{P}^b , Houtekamer and Mitchell (1998) first introduced a cutoff distance to limit the impact of observation on the state update beyond a certain distance. This is referred to as localization of the covariance. In the EnSRF, the covariance localization is achieved via multiplying the sample covariance (or regression coefficient) between an observation and a state variable on model grid by a localization factor that determined by tapering (or localization) function. The localization function is essentially a correlation function or a distance-dependent function with the value ranging from one to zero with an increasing distance. By using the covariance localization, the impact of a given observation on the state update can be limited around the observation location.

This study adopts the Gaspari and Cohn (GC) function (Gaspari & Cohn, 1999), which is widely used in atmospheric data assimilation and is denoted as τ here, to taper the ensemble-based covariance. The GC function is parameterized by a localization length scale, L , that determines a distance beyond which the correlation becomes zeros. In the GSI-EnSRF, the localization factor, ρ , is equal to a vertical localization function, ρ_v , multiplied by a horizontal localization function, and ρ_h , in the spherical Cartesian coordinates. The localization factor is given as

$$\rho = \rho_v \times \rho_h \quad (9)$$

$$\rho_v = \tau \left(\frac{r_v}{L_v} \right) \quad (10)$$

$$\rho_h = \tau \left(\frac{r_h}{L_h} \right) \quad (11)$$

where r_h is the horizontal distance between an observation location and a model grid point and r_v is the difference in log-scale pressure levels of an observation location and a given model grid level. In other words, r_v is an absolute altitude difference between an observation height and a model grid level given in terms of the scale height. Both vertical localization and horizontal localization functions are specified by the GC function with a vertical localization length scale, L_v , and a horizontal localization length scale, L_h , respectively.

It is difficult to define a location of the RO sTEC observation because sTEC is a nonlocal quantity. Therefore, the tangent point of each raypath is adopted as an observation location for the sake of implementing the covariance localization. A tangent point is the point along a given raypath that is the closest to the Earth under the straight-line propagation. In the F region of the ionosphere, electron densities around a tangent

point usually account for a large proportion of the electron densities integrated to sTEC. Since the RO raypath for a given sTEC traverses a large distance through the ionosphere and plasmasphere, sTEC observations contain information about the plasma densities over a large spatial model domain. It is important to note that the covariance between a given sTEC observation and model state variables is still inhomogeneous and anisotropic, even after the GC function is applied to taper the sample covariance to localize the impact of observations around the tangent point.

Generally speaking, the smaller the ensemble size, the higher the sampling errors. The covariance localization and inflation is used to rectify the issues that arise from spurious correlations due to the sampling errors. This paper will focus on the impact of both the covariance localization and ensemble size on quality of EnSRF assimilation analysis. By comparing results from a number of observing system simulation experiments (OSSEs), the most effective ensemble size and the length scales of the covariance localization to assimilate sTEC data into GIP/TIE-GCM using the EnSRF are determined in this study.

2.3. Observation Operator

Since the observations are usually not co-located with the model grid points and the observed variable is often different from the model state variable, the model state variables need to be converted to the observed variables by using a forward (observation) operator, \mathbf{H} . In this study, the observation is RO sTEC; hence, the sTEC value needs to be computed by integrating electron densities on the GIP/TIE-GCM grid along the RO raypath to obtain the predicted value sTEC by the model ($\mathbf{H}\mathbf{x}^p$). For given hypothetical positions of GPS, GLONASS, and FORMOSAT-7/COSMIC-2 low-inclination satellites, the RO raypath geometry can be determined. After that, this raypath is discretized into 20 km segments, and the electron density at the center of each segment is interpolated from electron densities on the model grid. Finally, the model-predicted sTEC value is set to the sum of integrated electron densities at each segment along the entire path. The interpolation scheme similar to the one in Yue, Schreiner, Kuo, et al. (2014) is adapted to cope with the irregular grid distributions of the GIP.

3. Data Assimilation Experiments

OSSEs are one of widely used approaches to evaluate the potential impact of given observing systems before they are developed or deployed (Hoffman & Atlas, 2016). In OSSEs, the synthetic observations are simulated from the "true" state provided by a numerical model, often referred to as the nature run (NR), with the expected coverage, resolution, and accuracy of observation systems. Using synthetically generated observation, usual data assimilation experiments are carried out. Please note that the model ensemble members used in the data assimilation experiments are different from NR. By verifying the data assimilation results against the NR, the impact of assimilating observations from a hypothetical observing system on specification and forecasting of the geophysical system can be assessed.

The specific purpose of OSSEs here is to assess the ability of FORMOSAT-7/COSMIC-2 observing system to improve the low-latitude and midlatitude ionospheric specification and forecasting. In this study, a number of OSSEs with different covariance localization length scales and ensemble sizes are conducted. All OSSEs are conducted under low solar activity, geomagnetically quiet, and solstice conditions, from 00:00 UT to 12:00 UT of 1 January. Synthetic sTEC observations are assimilated hourly into the GIP/TIE-GCM as described below.

3.1. Initialization of the Model Ensemble

The model ensemble is generated by perturbing three main model drivers: F107, HP, and CP, according to a Gaussian distribution with the mean value of F107, HP, and CP set to 120 Solar Flux Unit (SFU), 16 GW, and 45 kV, and with the standard deviation set to 15 SFU, 2 GW, and 10 kV, respectively. When drawing the ensemble samples of these drivers from the respective Gaussian distribution, we assume that the F107 index is independent of both HP and CP, but HP is correlated to CP. Figure 1 shows the histogram of the model driver ensembles along with the Gaussian distribution function from which the ensemble members are randomly drawn. In addition, the NR is executed by running the GIP/TIE-GCM under higher solar and geomagnetic conditions than those for the ensemble mean. Specifically, the plasma density distributions of NR are generated with the F107, HP, and CP values of 140 SFU, 18 GW, and 55 kV, respectively.

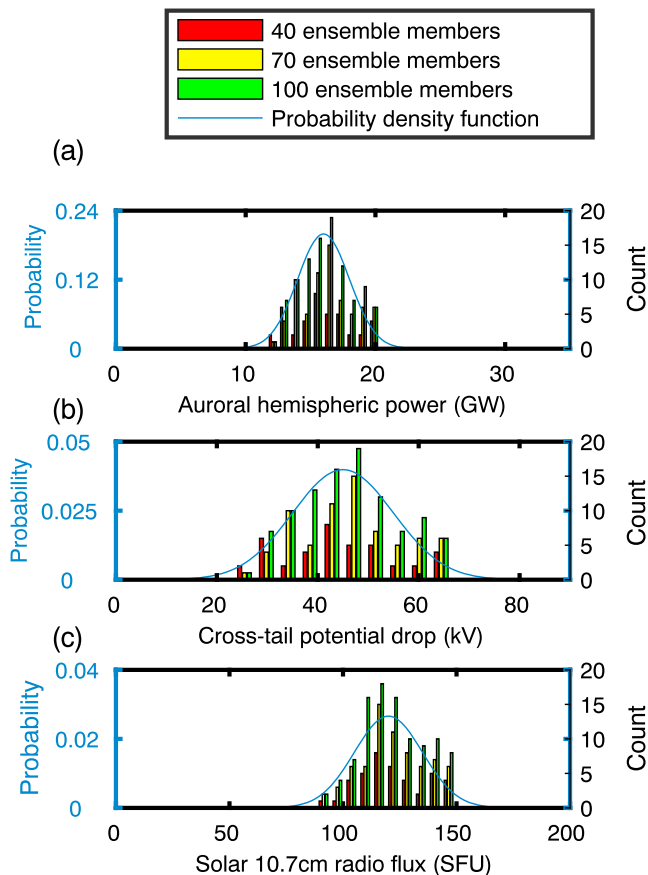


Figure 1. Histogram of (a) auroral hemispheric power, (b) cross-tail potential drop, and (c) solar 10.7 cm radio flux that used to drive the Global Ionosphere Plasmasphere (GIP)/Thermosphere Ionosphere Electrodynamic General Circulation Model (TIE-GCM) model ensemble, along with the underlying Gaussian probability distribution (shown in blue). The red, yellow, and green bars are the histograms of GIP/TIE-GCM drivers for the 40-, 70-, and 100-member ensembles, respectively.

ln(mb), are employed. For comparison, filtering experiments without the horizontal or/and vertical covariance localization are additionally executed. Considering the horizontal resolution of TIE-GCM is $5^\circ \times 5^\circ$ and an average raypath length that travels through the ionosphere is roughly 7000 km, the horizontal localization length scale for the GC function investigated here ranges from 500 to 10,000 km. The vertical localization length scales need to be given in terms of scale height in the GSI-EnSRF. Since the TIE-GCM vertical resolution is two levels per scale height and the total number of levels of TIE-GCM interface is 29, the range of vertical localization length scales for the GC function explored is from 0.5 to 7 scale height.

Because the vertical localization length scale is specified in terms of scale height, the covariance localization is anisotropic in terms of geometric height (km). The cut-off distance of the observation impact is farther away from an observation in the upward direction than the downward direction. For instance, the impact of observation located around the F2 peak is more aggressively localized in the bottomside ionosphere than in the topside ionosphere, which is favorable considering different physical mechanisms determining the *F* region and *E* region plasma density distributions.

According to the accuracy of TriG GNSS RO System, the observation error is 3 TEC unit for all data in OSSEs. Moreover, The weighting factor for covariance inflation in equation (8), w , is set to 0.9 based on experiments shown in Figure S1 in the supporting information, which compare two OSSEs with different values of weighting factor, $w = 0.1$ and $w = 0.9$. There is no significance difference, but an OSSE with a larger

Before data assimilation cycling, the model ensembles need to be spun up in order to allow enough time for each model ensemble member to reach the state that is dynamically balanced with perturbed drivers and to obtain the model ensemble with an adequate spread. In the spin-up period, the stand-alone TIE-GCM model is run for 23 days. After that, the thermospheric state obtained from a long integration of the stand-alone TIE-GCM is used as initial conditions to advance the GIP/TIE-GCM for another 5 days. Note that the NR is spun up in the same manner and that all model drivers are fixed during the spin-up and data assimilation cycling periods.

3.2. Synthetic FORMOSAT-7/COSMIC-2 Observations

The synthetic RO sTEC observations along the raypath between GPS and GLONASS satellites and FORMOSAT-7/COSMIC-2 low-inclination satellites are generated as follows. Using the same observation operator described above, the electron densities from NR on the model grid first interpolated the values along a raypath with 20 km segments, and then integrated over a raypath. After that, observational errors are added based on a centered Gaussian distribution with the standard deviation of 3 TEC unit. The sampling rate used for each RO event is 1 Hz. Roughly 300 to 400 FORMOSAT-7/COSMIC-2 RO events that amount to 200,000 sTEC data are assimilated into the GIP/TIE-GCM at each data assimilation cycle.

3.3. OSSE Design

The first set of OSSEs is conducted to determine the impact of the ensemble size on the EnSRF performance and the quality of assimilation analysis. The EnSRFs with three different model ensemble sizes 40, 70, and 100 are executed with the identical covariance localization setting. In the second set of OSSEs, the EnSRFs with an ensemble size selected based on the first set of OSSEs are run to further study the impact of covariance localization. In the covariance localization scheme, the GC functions with four different horizontal localization length scales, including 500, 1,000, 5,000, and 10,000 km, and four different vertical localization length scales, including 0.5, 1, 3, and 7

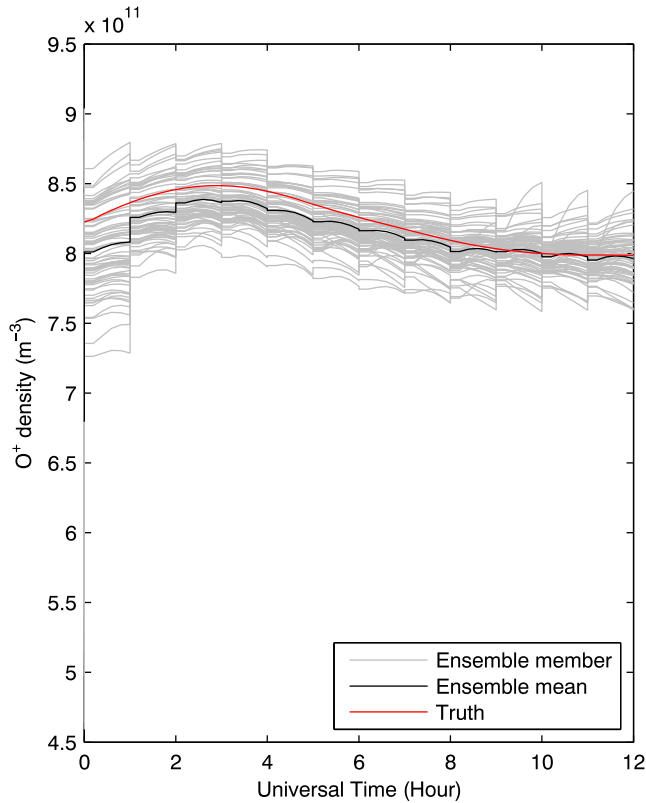


Figure 2. The atomic oxygen ion density averaged over the low- and midgeomagnetic latitudes from 200 to 500 km altitude is shown for each ensemble member from the observing system simulation experiment with the 70-member Ensemble Square Root Filter with a covariance localization with a length scale of 5,000 km in the horizontal direction. While the grey lines are for ensemble members, the red line is for the nature run and the black line is for the ensemble mean.

weighting factor performs slightly better. Therefore, a covariance inflation with $w = 0.9$ is applied in all the OSSEs presented in this study.

4. Results

The comparison of the prior and posterior ensemble distribution to the NR is presented in terms of the root-mean-square difference (RMSD) between the ensemble mean and NR of atomic oxygen ion density, computed over the geomagnetic low-latitude and midlatitude regions from 200 to 500 km altitudes, where FORMOSAT-7/COMSIC-2 low-inclination RO data have the greatest influence on the assimilation analysis. In addition to the OSSEs described in the previous section, a control ensemble forecast experiment is executed, with no data assimilation, using the same perturbed model driver parameters used to initialize the model ensemble for the OSSEs. The RMSD between the forecast ensemble mean and the NR is computed in the same manner as for the posterior and prior ensemble.

If the ionospheric data assimilation of sTEC by the EnSRF is successful, the RMSD should become smaller after the analysis step, suggesting a greater proximity of the estimated model state to the NR from which observations are sampled. The posterior ensemble spread is ought to become smaller than the prior ensemble spread, reflecting the uncertainty reduction in the state estimation. In the forecast step, the RMSD of OSSE is likely to increase toward the level of RMSD of control ensemble forecast experiment because model drivers are not altered by assimilation and the same perturbed model drivers are used in both sets of the ensemble simulations. The ensemble spread should also grow larger during the forecast step of the EnSRF to reflect an increased degree of uncertainty in the state estimation. Through successive applications of the analysis and forecast steps, the RMSD should overall continue to decrease. Figure 2 displays how the

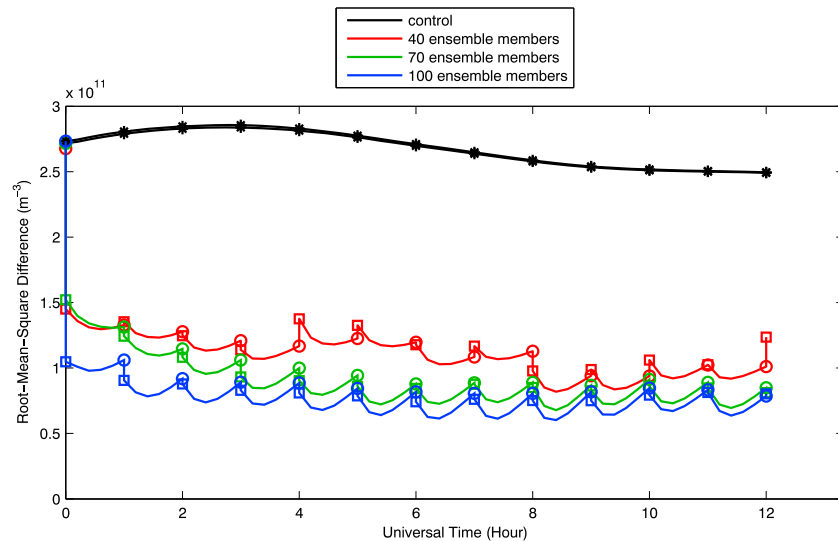


Figure 3. Root-mean-square difference (RMSD) between the ensemble mean and nature run of atomic oxygen ion density, computed over the midgeomagnetic and low-geomagnetic latitude regions from 200 to 500 km altitude, from 00:00 to 12:00 UT. Results from observing system simulation experiments with the 40-, 70-, and 100-member Ensemble Square Root Filter are shown in red, green, and blue, respectively. Covariance is localized with the Gaspari and Cohn function with 5,000 km length scale in the horizontal direction. No vertical covariance localization is applied. The circles and squares are the prior and posterior RMSD at analysis steps, respectively. The black line is for the RMSD of the 100-member control ensemble simulation.

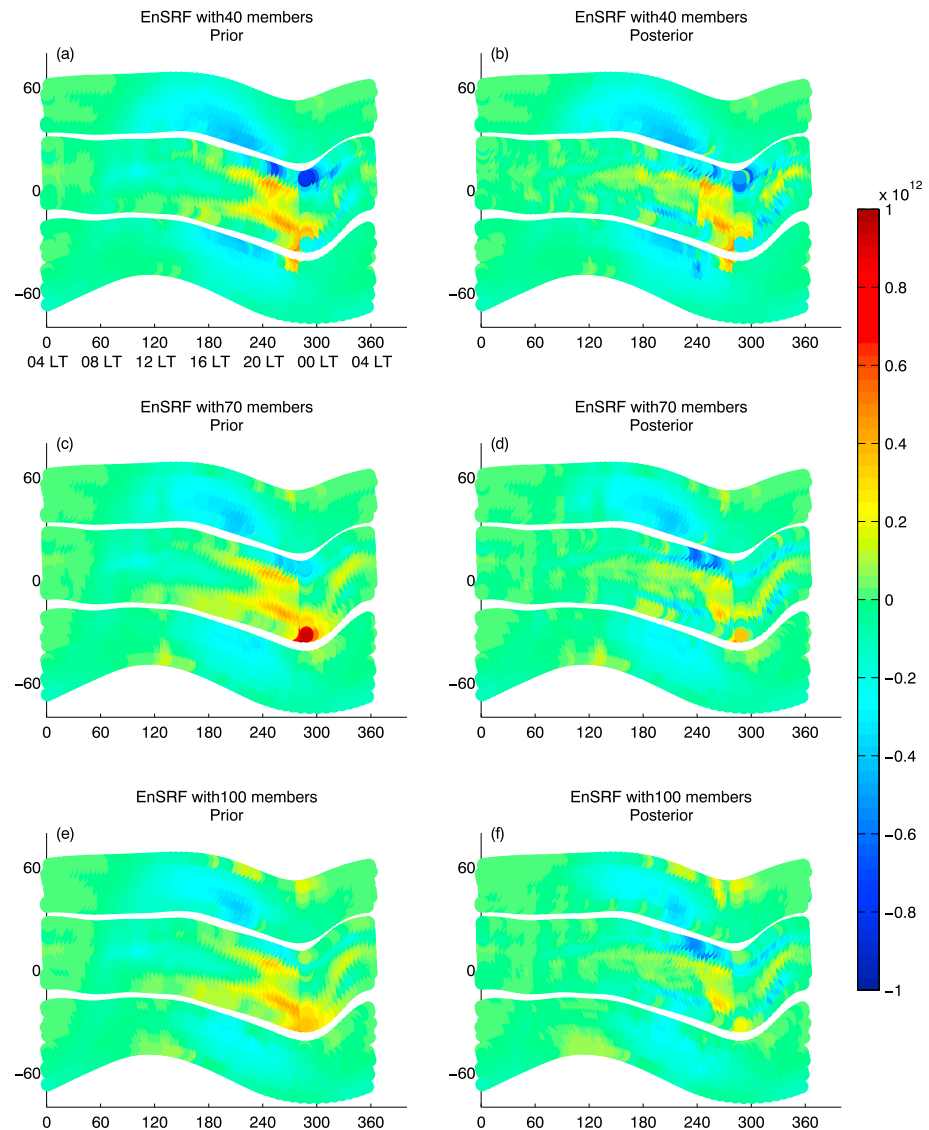


Figure 4. Differences of the atomic oxygen density between the prior and posterior ensemble mean and nature run at 330 km altitude at 04:00 UT are shown for the 40-, 70-, and 100-member Ensemble Square Root Filters (EnSRFs). Observing system simulation experiment results shown here are the same as those shown in Figure 3. (a) The prior mean bias for the 40-member EnSRF. (b) The posterior mean bias for the 40-member EnSRF. (c) The prior mean bias for the 70-member EnSRF. (d) The posterior bias for the 70-member EnSRF. (e) The prior mean with the 100-member EnSRF. (f) The posterior bias for the 100-member EnSRF.

ensemble mean and each ensemble member typically vary, as the global mean atomic oxygen density (in the geomagnetic low-latitude and midlatitude regions from 200 to 500 km altitudes, over the course of the entire data assimilation experiment). At the update step, all the ensemble members (grey lines) and the ensemble mean (black line) shift closer to the NR (red line), and the ensemble spread becomes smaller, representing the uncertainty reduction after incorporating the observation information into model ensemble. After that, the ensemble members diverge away from the NR and the ensemble spread grows larger during the forecast steps, representing the increasing of uncertainty.

4.1. Impact of Size of Model Ensemble on GSI Ionosphere Analysis

Figure 3 shows the RMSD of the OSSE results obtained from the EnSRF with 40, 70, and 100 ensemble members. In these experiments, the GC localization function is used to localize the covariance in the horizontal direction with a length scale of 5,000 km. No localization is applied in the vertical direction. The RMSD of

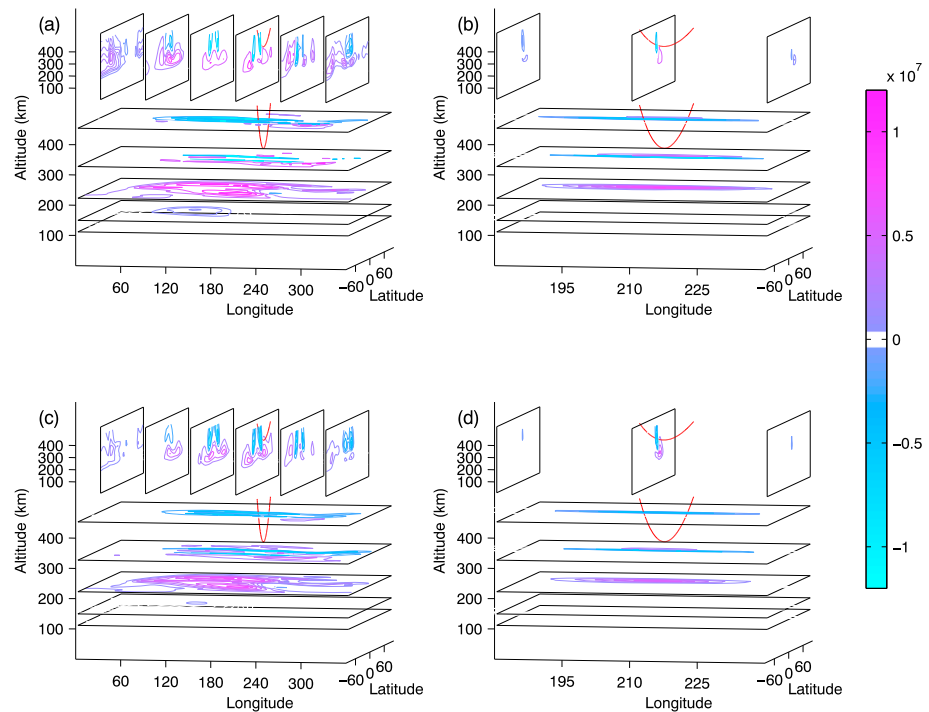


Figure 5. Examples of the prior covariance between a given slant total electron content (sTEC) observation (whose raypath is shown in red) and atomic oxygen ion density on model grid are shown for the 40- and 70-member EnSRF at 04:00 UT. The tangent point of this sTEC observation is located at 212° longitude and 22.3° latitude and at 04:00 UT. (a) The sample covariance estimated from 40 ensemble members. (b) The same covariance but with covariance localization applied in the horizontal direction. (c) The sample covariance estimated from 70 ensemble members. (d) The same covariance but with covariance localization applied in the horizontal direction. The localization scheme is used with the GC function with a length scale of 5,000 km.

the control ensemble forecast experiment with 100 ensemble members is also shown in Figure 3. Note that the performance of the control ensemble forecast experiment does not make much difference among the ensemble size of 40, 70, and 100.

The RMSD of these three OSSEs is generally smaller than that of control ensemble experiment over the entire data assimilation experiment of 12 hr. This indicates that the EnSRF can improve the ionospheric specification by bringing the model ensemble closer to the NR by assimilation of sTEC observations. The most significant improvement occurs in the first assimilation cycle. This is because the ionosphere of NR is biased to be higher in comparison to the ensemble mean, and the first update step is particularly effective in making a gross correction of the global atomic oxygen ion density distribution. This behavior is explored further with respect to a choice of the covariance localization parameters later. During the forecast step, the RMSD decreases for about 30 min and increases toward to the RMSD value of the control ensemble forecast experiment. This behavior will be future discussed in the next section.

As suggested by the RMSD, the performance of the EnSRF improves with an increasing number of ensemble members with the 100-member EnSRF at the best among three filters. The same conclusion holds for the comparison of 40-, 70-, and 100-member EnSRFs with different settings of covariance localization (see Figures S2, S3, and S4). Comparing with the 100-member EnSRF, the EnSRF with 70 ensemble members results in a larger RMSD at the beginning of data assimilation experiment, but the RMSD gradually reduces over time. At the end of the 12-hour data assimilation experiment, the ratio of RMSD to that of the control experiment is 0.3261 and 0.3215 for 70- and 100-member EnSRFs, respectively. The performance of these two filters is similar.

Unlike the 70- and 100-member EnSRFs, the 40-member EnSRF's performance is inconsistent, and some of the GIP/TIE-GCM ensemble simulations become numerically unstable during forecast steps. At the beginning, the behaviors of RMSD for the 40- and 70-member EnSRFs are similar, but the posterior

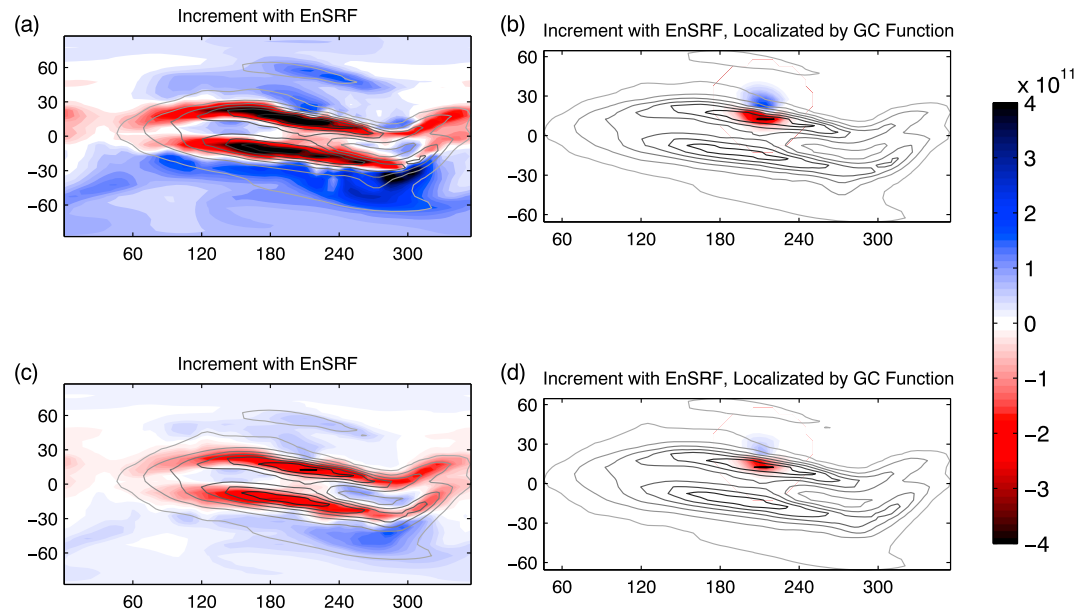


Figure 6. The analysis increment at 04:00 UT on a model pressure level corresponds to about 350 km in altitude for a slant total electron content (sTEC) observation whose tangent point is located at 212° longitude and 22.3° latitude estimated by 40- and 70-member Ensemble Square Root Filter (EnSRFs). The grey scale background contour lines represent the prior mean atomic oxygen ion density distribution. The darker colors represent higher densities. (a) The analysis increment estimated by the 40-member EnSRF. (b) The analysis increment estimated by the 40-member EnSRF with covariance localization. (c) The analysis increment estimated by the 70-member EnSRF. (d) The analysis increment estimated by the 70-member EnSRF with covariance localization.

RMSD for the 40-member EnSRF becomes larger than the prior RMSD after the fifth update step at 04:00 UT, which implies poor performance of the EnSRF. Under certain localization settings, even worse performance has been observed (see Figures S2, S3, and S4). In Figure 4, differences of the atomic oxygen density between the prior and posterior ensemble mean and NR at 330 km altitude at 04:00 UT are shown for the 40-, 70-, and 100-member EnSRF. Positive values indicate a positive bias, meaning that the atomic oxygen ion density of ensemble mean is larger than that of the NR, and vice versa for negative values. Since the error covariance estimated by the 40-member ensemble is not accurate enough, the posterior biases become larger than the prior biases in some regions. Although OSSEs with the 70- and 100-member EnSRFs also have some problems in low- and midlatitudes of postnoon and premidnight regions, the magnitude of biases decreases with an increasing ensemble size. As a result, when forecasting, the GIP/TIE-GCM is more stable if the model state is initialized by the EnSRF with the ensemble size of 70 or higher. At the end of whole data assimilation experiment at 12:00 UT, the ratio of RMSD to that of the control experiment is 0.5021 for the 40-member EnSRF, which is fairly large in comparison to the 70- and 100-member EnSRFs.

Figures 5a–5d display how the prior covariance between a given sTEC observation and the atomic oxygen densities on the model grid looks in the 40- and 70-member EnSRFs without and with the covariance localization. The raypath of this sTEC observation appears bended in these panels because it is displayed in the longitude-latitude-altitude coordinates. The tangent point of this raypath is located in the dayside EIA region at 350 km in altitude. Figure 6 shows the analysis increment, for the same cases, along with the prior mean atomic oxygen ion density distribution as the grey scale background contour. Note that the analysis increment refers to $\mathbf{K}(\bar{\mathbf{y}}^o - \mathbf{H}\bar{\mathbf{x}}^b)$ in equation (1). It is clear that, when no covariance localization is applied, the analysis increment obtained using the 70-member ensemble reflects better to the prior distribution in comparison to the 40-member case. As shown in Figure 6a, the analysis increment of the 40-member EnSRF, without covariance localization, is remarkably large even at distance far from the tangent point. When the covariance is estimated with a larger-size ensemble, these spuriously large covariance values are reduced and a more reasonable increment is obtained.

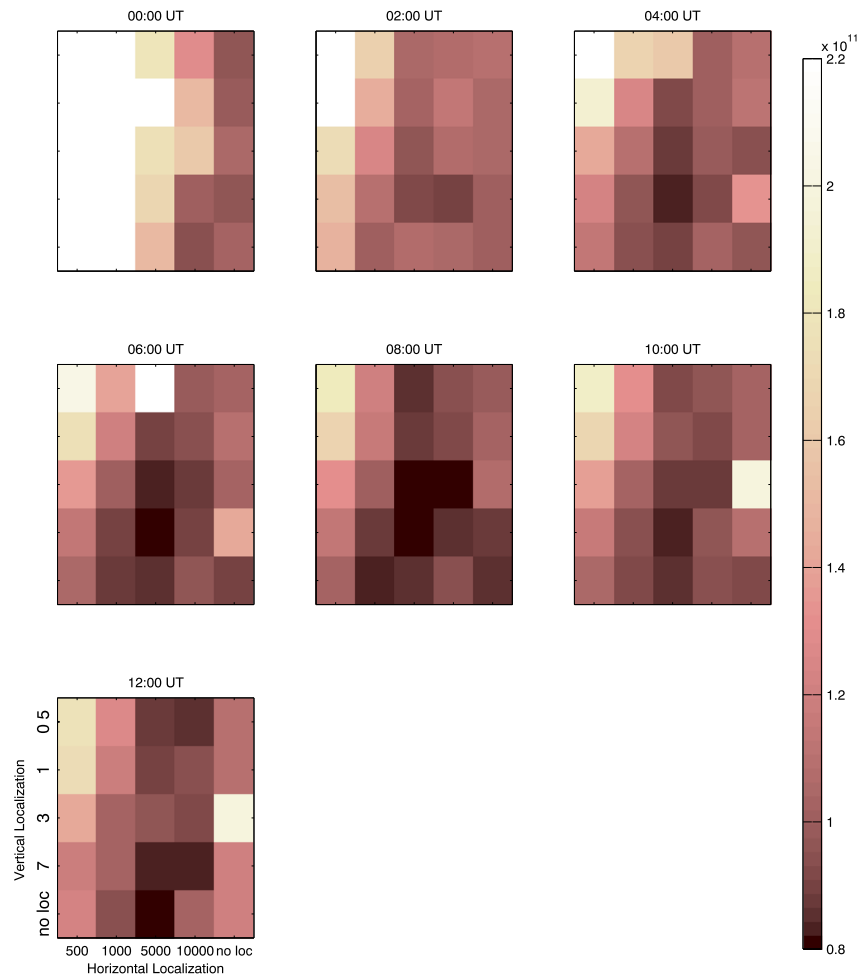


Figure 7. The posterior root-mean-square difference (RMSD), computed over the low- and midgeomagnetic latitude regions, from 200 to 500 km altitude, is shown for comparison of the EnSRF with different covariance localizations at 00:00, 02:00, 04:00, 06:00, 08:00, 10:00, and 12:00 UT. The Ensemble Square Root Filter with covariance localization using the Gaspari and Cohn function with four different vertical localization length scales (including 0.5, 1, 3, and 7 scale heights) and four different horizontal localization length scales (including 500, 1,000, 5,000, 10,000 km) are applied in the observing system simulation experiments. The posterior RMSD for the EnSRF without covariance localization is shown for comparison.

4.2. Impact of Covariance Localization on GSI Ionosphere Analysis

Figure 7 shows the RMSD of the posterior ensemble from OSSEs with the EnSRF with and without covariance localizations at 00:00, 02:00, 04:00, 06:00, 08:00, 10:00, and 12:00 UT. Note that the RMSD is computed over the same region as for Figure 3. For the vertical localization, the GC function with four different vertical localization length scales, including 0.5, 1, 3, and 7 scale heights, is applied. For the horizontal localization, the GC function with four different horizontal localization length scales, including 500, 1,000, 5,000, and 10,000 km, is applied. The ensemble size is 70 for all OSSEs shown here. The RMSD over entire data assimilation cycles can be found in Figures S5–S9. The RMSD of OSSEs that uses the GC function with smallest horizontal and vertical localization length scales in the localization scheme is considerably larger than other OSSEs. This suggests the need of careful tuning of covariance localization parameter.

The RMSD is reduced dramatically at the first update step, especially if the covariance is not localized or localized with the GC function with a large length scale in both the horizontal and vertical directions. As mentioned earlier, this is because the gross correction of the prior ensemble, here biased to be higher, is more effective with no localization of the covariance. In comparison, for the same set of observations, such a reduction is less dramatic for OSSEs with covariance localization with the GC function with a smaller length scale, but there is a steady reduction of RMSD over many assimilating cycles. At the end of data assimilation

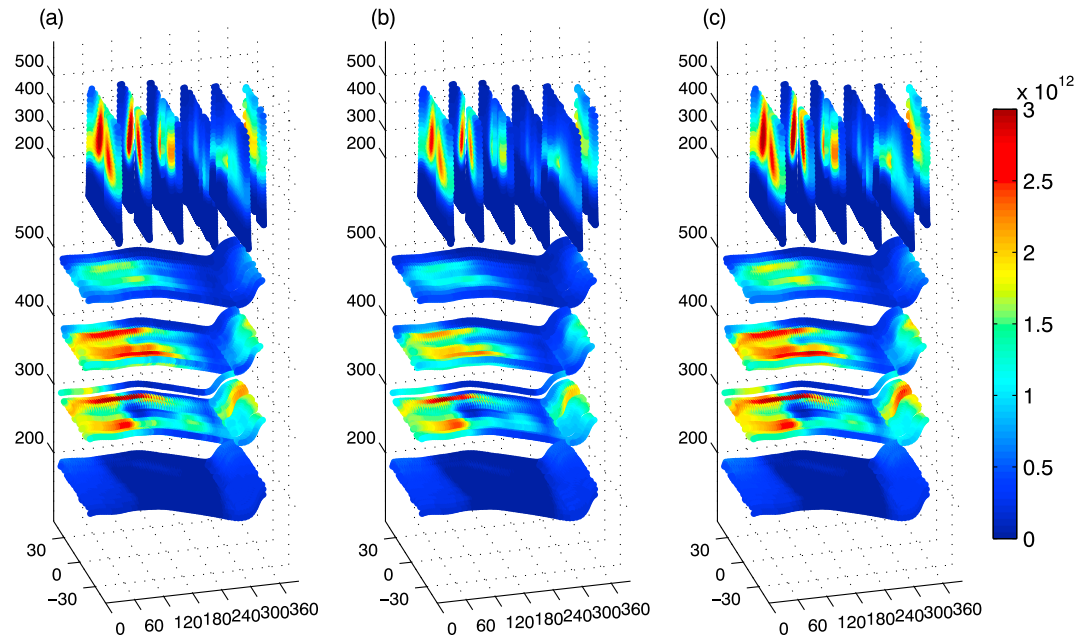


Figure 8. The atomic oxygen ion density distribution at 12:00 UT. (a) The posterior mean from the observing system simulation experiment with the 70-ensemble Ensemble Square Root Filter. (b) The mean of the ensemble control experiment. (c) The nature run.

experiment at 12:00 UT, the EnSRF with covariance localization with a length scale of 5,000 km in the horizontal direction and with no vertical localization leads to the smallest RMSD.

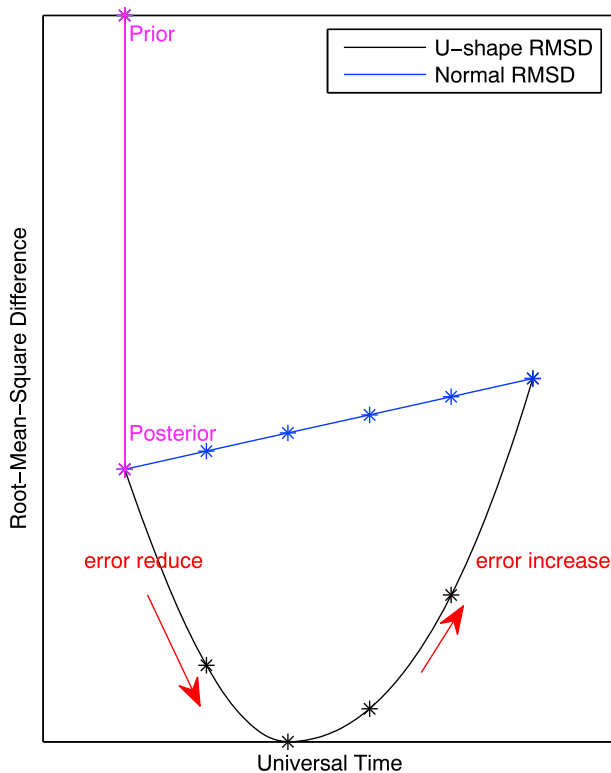


Figure 9. Illustration of the U-shape root-mean-square difference (RMSD; black line) and the RMSD from a typical experiment (blue line).

A choice of the localization length scale in the horizontal direction affects the assimilation analysis considerably as suggested by the RMSD magnitude. The EnSRF with localization with the GC function with a length scale of 5,000 to 10,000 km in the horizontal direction leads to the smallest error regardless of a choice of vertical localization length scale. Comparing with the horizontal direction, the impact of a vertical localization length scale appears to be minor. In general, the use of larger vertical localization length scales results in a smaller RMSD.

As shown in Figure 4, larger analysis biases in regions such as the low- and midlatitudes of postnoon and premidnight regions still need to be reduced by applying the covariance localization with a certain length scale. To further improve specification of the ionosphere in the EIA and boundary of high- and midlatitude regions, a localization function that is estimated specifically for the sTEC data using a method proposed by Anderson and Lei (2013) might be helpful in the future.

In summary, the second set of OSSEs demonstrates that, if the GC function is used to localize the covariance in the EnSRF, the most appropriate range of the horizontal localization length scale for sTEC data assimilation is from 5,000 km to 10,000 km. No vertical localization appears to be the most effective. Figure 8 shows the atomic oxygen ion density at the end of OSSE at 12:00 UT along with the mean of the control ensemble simulation and the NR. In this OSSE, the 70-member EnSRF is used with the GC function with a length scale of 5,000 km to localize the impact of observation in the horizontal direction. A visual inspection of these panels shows that the assimilation analysis shown in Figure 8a is closer to the NR shown in Figure 8c in comparison to the control simulation shown in Figure 8b.

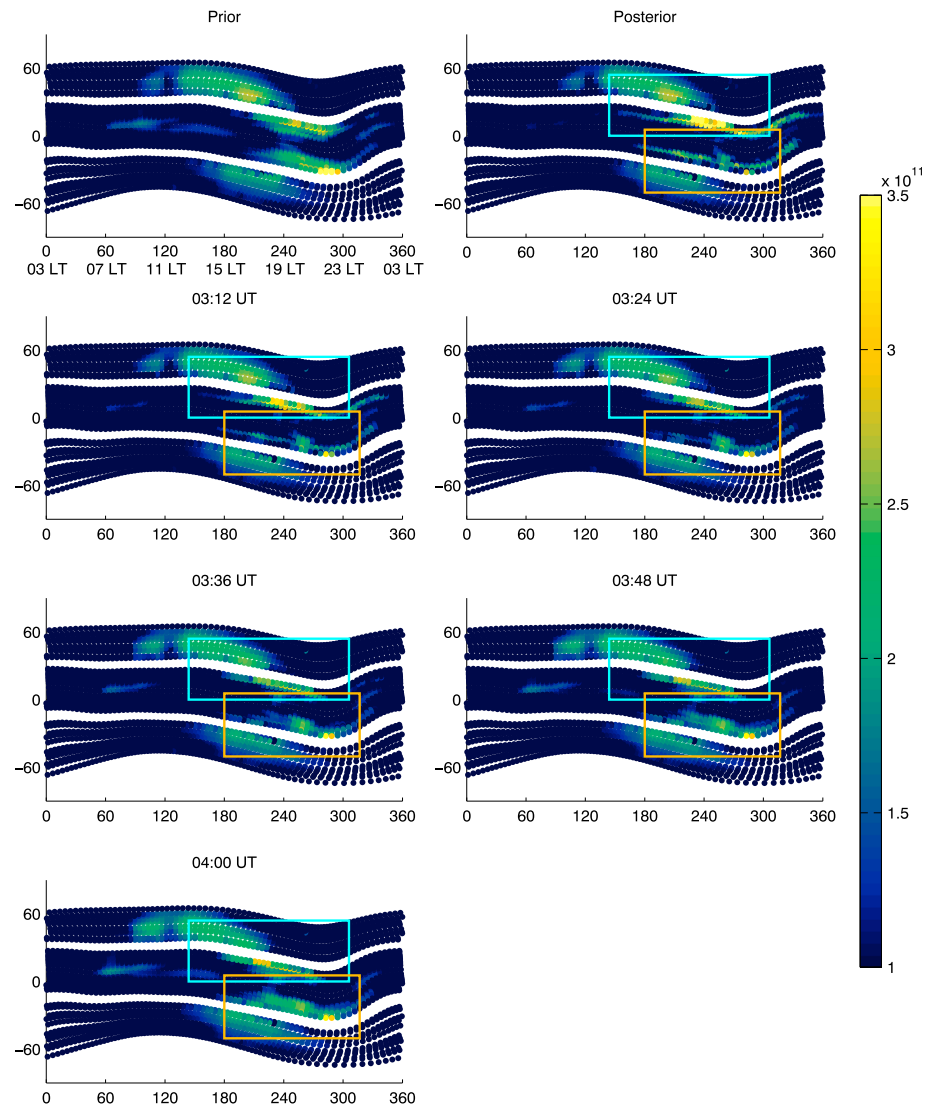


Figure 10. The root-mean-square difference (RMSD) computed for each fluxtube, from the difference between the atomic oxygen ion density of ensemble mean and nature run from 200 to 500 km altitude, during the fourth data assimilation cycle from 03:00 to 04:00 UT. This is the same observing system simulation experiment as presented in Figure 3 and executed with the 70-ensemble Ensemble Square Root Filter. Orange and cyan boxes denote regions A and B. (top row) RMSD before and after the assimilation update at 03:00 UT. The other panels are the RMSD during the forecast with a 12 min interval.

5. Discussion

As shown in Figure 3, during almost all forecast steps of the EnSRFs, the RMSD grows smaller for about 30 min before starting to grow larger as expected. This peculiar behavior, here referred to as the “U-shape” RMSD, suggests that the GIP/TIE-GCM ensemble mean continues shifting toward the NR whose driver setting is slightly higher than the ensemble mean setting during the forecast step. As illustrated in Figure 9, the RMSD is expected to continuously grow during a forecast step.

To understand this better, the model state is examined in detail during the forecast step of the fourth data assimilation cycle when the U-shape RMSD is the most evident from 03:00 to 04:00 UT (see Figure 3). Figure 10 shows the RMSD computed along each magnetic field line from 200 to 500 km altitude at every 12 min from 03:00 to 04:00 UT. The OSSE used to compute these RMSD maps was obtained with the 70-member EnSRF. The high RMSD region appears roughly from 180° to 300° longitude that corresponds to 15:00 to 23:00 LT (postnoon to premidnight) at 03:00 UT. An apparently large RMSD region in the midgeomagnetic latitude at around 200° longitude (marked by a cyan box in Figure 10) and another

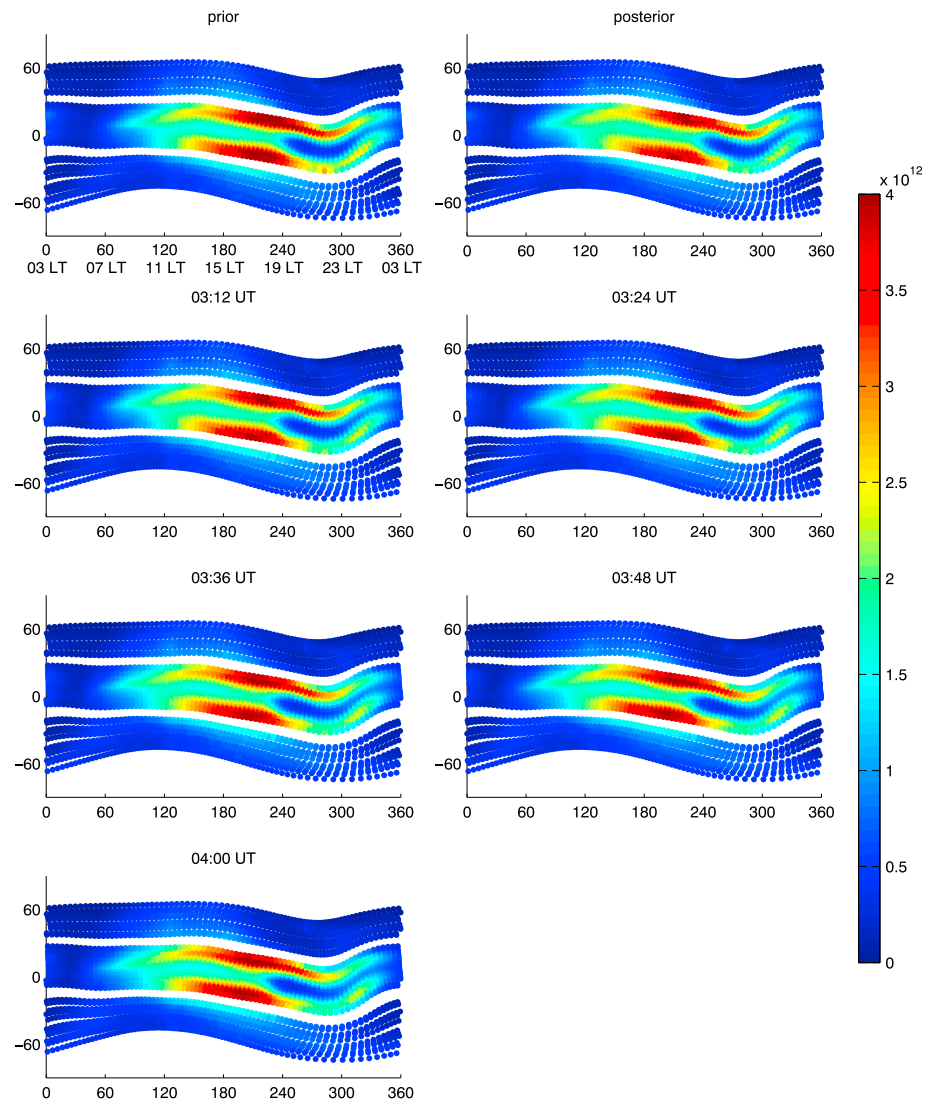


Figure 11. Horizontal distribution of the ensemble mean atomic oxygen ion density at 330 km altitude from 03:00 to 04:00 UT, in the same order as Figure 10.

large RMSD region in low-geomagnetic latitude at around 300° longitude (marked by an orange box) are referred to as regions A and B, respectively. In the course of the forecast step, the RMSD in the region A becomes smaller, while the RMSD in the region B increases. The error reduction and increase in these regions are responsible for the U-shape RMSD computed over a large model domain.

Figure 11 shows the distribution of atomic oxygen ion density at 330 km altitude, and the difference from the NR is shown in Figure 12 in the same format as in Figure 4 where the positive values mean that the ensemble mean is larger than the NR. Before the assimilation update at 03:00 UT, the ensemble mean is significantly larger than the NR in the postnoon to premidnight area in the low-geomagnetic latitude. The EnSRF corrects the density globally but overcorrects in the postnoon to premidnight area as indicated by negative values in the posterior bias map shown in Figure 12. Over the course of the forecast step, these negative biases become smaller, bringing the midgeomagnetic latitude ionosphere state closer to the NR, while the positive biases start appearing again. The regions of positive and negative biases agree with the regions with large errors in Figure 10. At the end of current data assimilation cycle, the ensemble mean again becomes larger than that of the NR.

From the postnoon to premidnight, the photoionization production process becomes weaker and the loss process through recombination becomes more dominant. As shown in Figure 12, the overall atomic

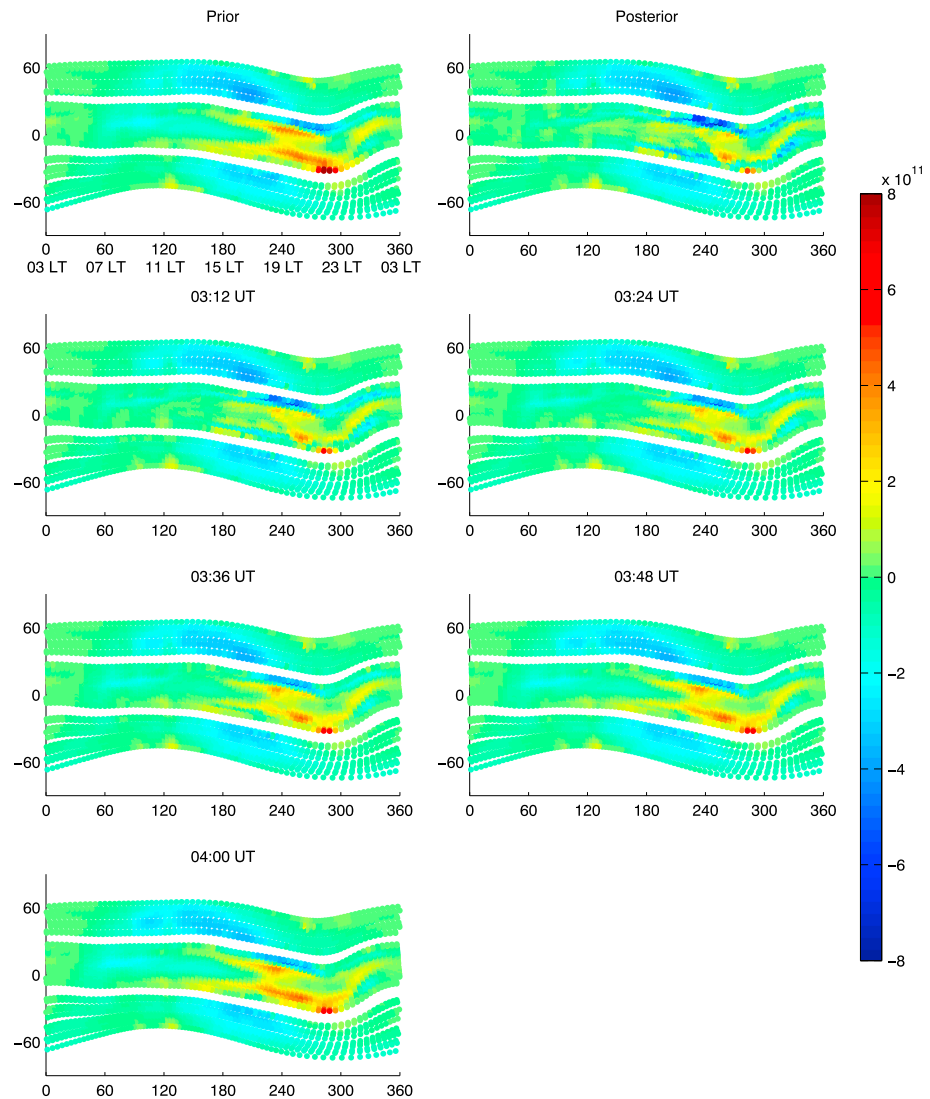


Figure 12. Differences of the atomic oxygen ion density between the ensemble mean and nature run, in the same order as in Figure 10.

oxygen ion densities in the OSSE are smaller than that in the NR at the beginning of forecast step and become larger. This implies that the loss rate of atomic oxygen ion density in the OSSE is slower than that of the NR.

Because synthetic sTEC observations are sampled from the NR with a higher level of the solar EUV flux, the *F* region peak density in the OSSE is being place in the atmosphere with less molecular concentration in comparison to the NR when the EnSRF brings both the peak density and peak height up. This results in a smaller loss rate of the atomic oxygen ion of OSSE than that of the NR through recombination with molecular species and leads to a positive bias during postnoon and premidnight that appears at the end of forecast step. Examining the OSSE results shown in Figures 10, 11, and 12 in more detail, 62% of the magnetic fluxtubes located from 180° to 300° longitude experience an increase in both peak density and peak height by the assimilation update at 03:00 UT. Figure 13 shows the locations of foot points of these fluxtubes, which largely overlap with regions A and B. The loss rate of OSSE in the *F* region is smaller than that of the NR in the majority (83%) of those fluxtubes whose peak density height is corrected to higher.

In summary, the U-shape RMSD results from limitations of the assimilation method rather than the intrinsic dynamical behaviors of the thermosphere and ionosphere. The analysis update of the atomic oxygen ion density in the local postnoon and premidnight regions is inadequate, resulting in a negative bias from the

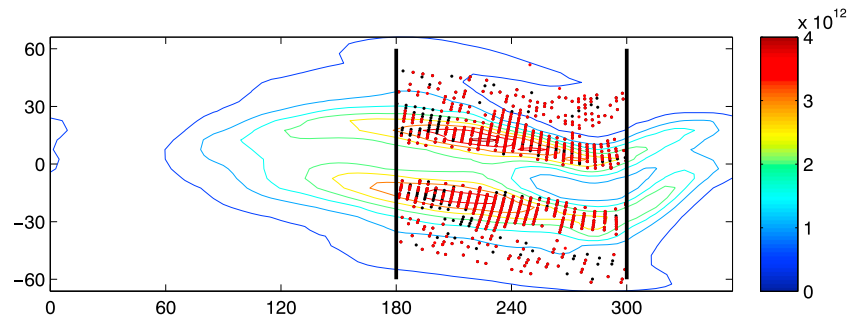


Figure 13. Foot points of the fluxtubes, over 180° and 300° longitude, whose peak density and peak height increase by the assimilation update at 03:00 UT. The black and red dots represent the foot points of fluxtubes in which the loss rate of observing system simulation experiment is larger and smaller than that of the NR, respectively. The background color contour represents the posterior mean distribution at 330 km altitude.

NR as shown in Figure 12. Because the neutral composition is unaffected by the assimilation update, the EnSRF brings the peak density larger and brings peak height to the atmospheric region with a less abundance of molecular species. As a result, the situation that the loss rate in the OSSE is smaller than that in the NR. These limitations should be overcome in the future, by updating the thermospheric compositions in the analysis step as has been done in Hsu et al. (2014) and by improving quality of assimilation analysis with the help of a nonparametric covariance localization function estimated for a specific observing system (e.g., Anderson & Lei, 2013).

A few issues with the current localization scheme need to be addressed in the future study. First, the bending angle of raypath that travel through the ionosphere and plasmasphere is very small, so we could consider the raypath as a straight line between GNSS (GPS and GLONASS) satellites and FORMOSAT-7/COSMIC-2 low-inclination satellites in normal Cartesian coordinate. It is ideal to adaptively localize the covariance along the raypath of a given sTEC observations. On the other hand, currently in the GSI, the horizontal distance is computed in the spherical Cartesian coordinate with precludes a true representation of distance. This discrepancy causes an incorrect adjustment in covariance localization. Second, the raypath travels over a large horizontal distance in the ionosphere but confined vertically. Although the vertical localization is expected to improve quality of the assimilation analysis, an aggressive vertical localization in the OSSEs results in discontinuous ion/electron density profiles that in turn introduce an undesirable unbalance in dynamical and chemical processes in the forecast step. A more comprehensible investigation of the covariance localization scheme for ensemble data assimilation of sTEC observation is needed to solve issues addressed above in the future.

6. Conclusions

Data assimilation is a powerful technique that can be used not only for monitoring the ionospheric weather but also for gaining a better understanding of various ionospheric phenomena. By systematically contrasting various observations and a model through the process of data assimilation, we are able to identify our lack of understanding of fundamental physical processes described in the first-principle model. Although this study focuses on assimilating the FORMOSAT-7/COSMIC-2 low-inclination RO sTEC data into the GSI ionosphere system with the practical aim of improving the ionospheric specification and forecasting, this technique will also be helpful for addressing science questions, for instance, regarding day-to-day variability of ionosphere by providing an instantaneous global picture of the ionosphere.

The GSI ionosphere is an ionospheric data assimilation system that is constructed with the NOAA GSI-EnSRF and GIP/TIE-GCM. The impact of sTEC on the low- and midlatitude ionosphere specification has been investigated through a comparative analysis of OSSEs. By using the GIP/TIE-GCM in conjunction with the EnSRF, the data assimilation analysis is produced with the benefit of a self-consistent coupling of the ionosphere and plasmasphere with the thermosphere in the forecast steps. The EnSRF is an ensemble-based data assimilation scheme, and detrimental effects of the sampling errors caused by the use of a finite number of ensemble need to be rectified in order to construct a stable and effective filtering system and to yield high quality

data assimilation analysis. A number of the OSSEs are carried out, with different ensemble sizes and different covariance localization scales, to examine the most suitable EnSRF parameters for sTEC data assimilation with the GIP/TIE-GCM.

Primary findings are summarized as follows:

1. Generally, data assimilation of FORMOSAT-7/COSMIC-2 RO data can improve the low- and midlatitude ionospheric specification. After the course of data assimilation cycles over 12 hr, the low- and midlatitude atomic oxygen ion density distribution of OSSEs becomes closer to that of the NR, which results in about 68% reduction of the RMSD in comparison to the control ensemble simulation.
2. For a given localization length scale, the EnSRF with a larger ensemble size (>70) consistently performs better for assimilation of RO sTEC. The use of the EnSRF with at least 70 ensemble members for sTEC data assimilation is recommended for future studies.
3. The RO sTEC data assimilation with the EnSRF is sensitive to a choice of the horizontal localization length scales. The covariance localization with the GC function with a length scale of 5,000–10,000 km in the horizontal direction helps in improving the stability of filtering and the quality of data assimilation analysis. On the other hand, the vertical localization appears to have a minor or unclear effect.

In the future, the EnSRF performance can be improved further by taking the following measures. Nonparametric localization functions that are designed specifically to sTEC data with a consideration of the RO raypath geometry, instead of a parametric function such as the GC, are desirable. In addition, updating the thermospheric compositions during the analysis step is considered essential to extend the utility of the FORMOSAT-7/COSMIC-2 RO data to further improve the ionospheric specification using the GSI-EnSRF.

Acknowledgments

This study is supported by the NOAA Space Weather Prediction Center and by the following grants: Taiwan Ministry of Science and Technology grant MOST 105-2119-M-008-020, National Space Organization grant NSPO-S-104083, NASA award NNX14AI17G, and AFOSR grant FA9550-15-1-0308. The authors thank NCAR COSMIC office for their great help with FORMOSAT-7/COSMIC-2 RO data. The authors would like to acknowledge the high-performance computing resource and support provided on Yellowstone by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation (ark:/85065/d7wd3xhc). We are also grateful for helpful guidance from Jeff Whitaker and Lili Lei. This work is presented at the 2017 International Team meeting for Ionospheric Space Weather Studied by RO and Ground-based GPS TEC Observations, which is led by Jann-Yenq Liu and supported by International Space Science Institute, Bern, Switzerland. The main data assimilation result presented in this paper is publicly available from <https://doi.org/10.17605/OSF.IO/S6UY3> (Hsu, 2017).

References

- Anderson, J., & Lei, L. (2013). Empirical localization of observation impact in ensemble Kalman filters. *Monthly Weather Review*, *141*(11), 4140–4153. <https://doi.org/10.1175/MWR-D-12-00330.1>
- Anderson, J. L. (2001). An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, *129*(12), 2884–2903. [https://doi.org/10.1175/1520-0493\(2001\)129%3C2884:AEAKFF%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129%3C2884:AEAKFF%3E2.0.CO;2)
- Anderson, J. L. (2003). A local least squares framework for ensemble filtering. *Monthly Weather Review*, *131*(4), 634–642. [https://doi.org/10.1175/1520-0493\(2003\)131%3C0634:Allsf%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(2003)131%3C0634:Allsf%3E2.0.CO;2)
- Anderson, J. L., & Anderson, S. L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, *127*(12), 2741–2758. [https://doi.org/10.1175/1520-0493\(1999\)127%3C2741:AMCIOT%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127%3C2741:AMCIOT%3E2.0.CO;2)
- Chartier, A. T., Matsuo, T., Anderson, J. L., Collins, N., Hoar, T. J., Lu, G., et al. (2016). Ionospheric data assimilation and forecasting during storms. *Journal of Geophysical Research: Space Physics*, *121*, 764–778. <https://doi.org/10.1002/2014JA020799>
- Chen, C. H., Lin, C. H., Matsuo, T., Chen, W. H., Lee, I. T., Liu, J. Y., et al. (2016). Ionospheric data assimilation with thermosphere-ionosphere-electrodynamics General Circulation Model and GPS-TEC during geomagnetic storm conditions. *Journal of Geophysical Research: Space Physics*, *121*, 5708–5722. <https://doi.org/10.1002/2015JA021787>
- Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *Journal of Geophysical Research*, *99*, 10,143–10,162. <https://doi.org/10.1029/94JC00572>
- Gaspari, G., & Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 723–757. <https://doi.org/10.1002/qj.49712555417>
- Hajj, G. A., Wilson, B. D., Wang, C., Pi, X., & Rosen, I. G. (2004). Data assimilation of ground GPS total electron content into a physics-based ionospheric model by use of the Kalman filter. *Radio Science*, *39*, RS1505. <https://doi.org/10.1029/2002RS002859>
- Hamill, T. M., Whitaker, J. S., & Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, *129*(11), 2776–2790. [https://doi.org/10.1175/1520-0493\(2001\)129%3C2776:DDFOBE%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129%3C2776:DDFOBE%3E2.0.CO;2)
- Hoffman, R. N., & Atlas, R. (2016). Future observing system simulation experiments. *Bulletin of the American Meteorological Society*, *97*(9), 1601–1616. <https://doi.org/10.1175/BAMS-D-15-00200.1>
- Houtekamer, P. L., & Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, *126*(3), 796–811. [https://doi.org/10.1175/1520-0493\(1998\)126%3C0796:DAUAEK%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126%3C0796:DAUAEK%3E2.0.CO;2)
- Houtekamer, P. L., & Mitchell, H. L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *129*(1), 123–137. [https://doi.org/10.1175/1520-0493\(2001\)129%3C0123:ASEKFF%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129%3C0123:ASEKFF%3E2.0.CO;2)
- Hsu, C. T. (2017). Assessment of the impact of FORMOSAT-7/COSMIC-2 GNSS RO observations on mid- and low-latitude ionosphere specification: Observing system simulation experiments using Ensemble Square Root Filter. Retrieved from Open Science Framework. <https://doi.org/10.17605/OSF.IO/S6UY3>
- Hsu, C. T., Matsuo, T., Wang, W. B., & Liu, J. Y. (2014). Effects of inferring unobserved thermospheric and ionospheric state variables by using an ensemble Kalman filter on global ionospheric specification and forecasting. *Journal of Geophysical Research: Space Physics*, *119*, 9256–9267. <https://doi.org/10.1002/2014JA020390>
- Ide, K., Courtier, P., Ghil, M., & Lorenc, A. C. (1997). Unified notation for data assimilation: Operational, sequential and variational. *Journal of the Meteorological Society of Japan. Ser. II*, *75*(1B), 181–189. https://doi.org/10.2151/jmsj1965.75.1B_181
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45. <https://doi.org/10.1115/1.3662552>
- Komjathy, A., Wilson, B., Pi, X., Akopian, V., Dumett, M., Iijima, B., et al. (2010). JPL/USC GAIM: On the impact of using COSMIC and ground-based GPS measurements to estimate ionospheric parameters. *Journal of Geophysical Research*, *115*, A02307. <https://doi.org/10.1029/2009JA014420>

- Lee, I. T., Matsuo, T., Richmond, A. D., Liu, J. Y., Wang, W., Lin, C. H., et al. (2012). Assimilation of FORMOSAT-3/COSMIC electron density profiles into a coupled thermosphere/ionosphere model using ensemble Kalman filtering. *Journal of Geophysical Research*, *117*, A10318. <https://doi.org/10.1029/2012JA017700>
- Matsuo, T., & Araujo-Pradere, E. A. (2011). Role of thermosphere-ionosphere coupling in a global ionospheric specification. *Radio Science*, *46*, RS0D23. <https://doi.org/10.1029/2010RS004576>
- Matsuo, T., Lee, I. T., & Anderson, J. L. (2013). Thermospheric mass density specification using an ensemble Kalman filter. *Journal of Geophysical Research: Space Physics*, *118*, 1339–1350. <https://doi.org/10.1002/jgra.50162>
- Millward, G. H., Müller-Wodarg, I. C. F., Aylward, A. D., Fuller-Rowell, T. J., Richmond, A. D., & Moffett, R. J. (2001). An investigation into the influence of tidal forcing on *F* region equatorial vertical ion drift using a global ionosphere-thermosphere model with coupled electrodynamics. *Journal of Geophysical Research*, *106*, 24,733–24,744. <https://doi.org/10.1029/2000JA000342>
- Pedatella, N. M., Forbes, J. M., Maute, A., Richmond, A. D., Fang, T. W., Larson, K. M., & Millward, G. (2011). Longitudinal variations in the *F* region ionosphere and the topside ionosphere-plasmasphere: Observations and model simulations. *Journal of Geophysical Research*, *116*, A12309. <https://doi.org/10.1029/2011JA016600>
- Richmond, A. D. (1995). Ionospheric electrodynamics using magnetic apex coordinates. *Journal of Geomagnetism and Geoelectricity*, *47*(2), 191–212. <https://doi.org/10.5636/jgg.47.191>
- Richmond, A. D., Ridley, E. C., & Roble, R. G. (1992). A thermosphere/ionosphere General Circulation Model with coupled electrodynamics. *Geophysical Research Letters*, *19*, 601–604. <https://doi.org/10.1029/92GL00401>
- Scherliess, L., Schunk, R. W., Sojka, J. J., Thompson, D. C., & Zhu, L. (2006). Utah State University Global Assimilation of Ionospheric Measurements Gauss-Markov Kalman filter model of the ionosphere: Model description and validation. *Journal of Geophysical Research*, *111*, A11315. <https://doi.org/10.1029/2006JA011712>
- Scherliess, L., Thompson, D. C., & Schunk, R. W. (2009). Ionospheric dynamics and drivers obtained from a physics-based data assimilation model. *Radio Science*, *44*, RS0A32. <https://doi.org/10.1029/2008RS004068>
- Schunk, R. W., Scherliess, L., Eccles, V., Gardner, L. C., Sojka, J. J., Zhu, L., et al. (2016). Space weather forecasting with a Multimodel Ensemble Prediction System (MEPS). *Radio Science*, *51*, 1157–1165. <https://doi.org/10.1002/2015RS005888>
- Schunk, R. W., Scherliess, L., Sojka, J. J., Thompson, D. C., Anderson, D. N., Codrescu, M., et al. (2004). Global Assimilation of Ionospheric Measurements (GAIM). *Radio Science*, *39*, RS1502. <https://doi.org/10.1029/2002RS002794>
- Wang, C., Hajj, G., Pi, X., Rosen, I. G., & Wilson, B. (2004). Development of the Global Assimilative Ionospheric Model. *Radio Science*, *39*, RS1506. <https://doi.org/10.1029/2002RS002854>
- Whitaker, J. S., & Hamill, T. M. (2002). Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, *130*(7), 1913–1924. [https://doi.org/10.1175/1520-0493\(2002\)130%3C1913:EDAWPO%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130%3C1913:EDAWPO%3E2.0.CO;2)
- Whitaker, J. S., & Hamill, T. M. (2012). Evaluation methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, *140*(9), 3078–3089. <https://doi.org/10.1175/MWR-D-11-00276.1>
- Yue, X., Schreiner, W. S., Kuo, Y.-H., Braun, J. J., Lin, Y.-C., & Wan, W. (2014). Observing system simulation experiment study on imaging the ionosphere by assimilating observations from ground GNSS, LEO-based radio occultation and ocean reflection, and cross link. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(7), 3759–3773. <https://doi.org/10.1109/TGRS.2013.2275753>
- Yue, X., Schreiner, W. S., Pedatella, N., Anthes, R. A., Mannucci, A. J., Straus, P. R., & Liu, J.-Y. (2014). Space weather observations by GNSS radio occultation: From FORMOSAT-3/COSMIC to FORMOSAT-7/COSMIC-2. *Space Weather*, *12*, 616–621. <https://doi.org/10.1002/2014SW001133>