Referee #2

This study is related to the data assimilation of AMSE-A microwave radiance data (and additional PrepBUFR data) by DART (configured with EAKF) coupled with CESM. The AMSU-A observations were bias-corrected, and the observation errors were estimated. After these preprocessing steps, the observations were assimilated, and the results were validated and discussed. The study provided some interesting results, and the discussions were thought-provoking. Nevertheless, I have some concerns which were summarized below.

We sincerely appreciate your comments and concerns about this manuscript. Detailed responses to your comments are precisely described below. In this response letter, to separate the referee's comments from authors' responses, your original comments are presented in an Italic font with an underline.

1. L215-216: Do you mean that extra experiments were performed to determine the optimal spatial thinning length? If so, could you please provide some details about how the experiments were designed in order to estimate the optimal thinning length?

[Reply]
As you mentioned, extra assimilation experiments were conducted using the same DART assimilation system, in order to decide the suitable spatial thinning distance between different AMSU-A pixels. To make this point clear, we revised some sentences as follow:
[Old, lines 215-216]
"In this study, the AMSU-A observations are spatially thinned at an interval of about 290 km that was empirically estimated with multiple pre-trial runs."
[New]
"To choose the optimal spatial thinning distance, we performed four extra assimilation runs in which different spatial thinning distance (i.e., 96 km, 192 km, 288 km and 384 km) was applied. These distances are multiples of the AMSU-A FOV footprint size (~48 km in nadir). The thinning interval of 288 km resulted in the largest analysis impact, so that distance was used to thin the observations in this study."

2. L221-223: I have two comments here.
1) Could you please provide some references to this method (i.e., to estimate the bias by averaging the departures between the observed and simulated radiances)? Perhaps the study by Scheck et al. (2018) should be cited (section 5, P677). Although Scheck et al. (2018) focus on the visible imagery, their method should be applicable to the microwave imagery.
[Scheck, L., Weissmann, M., and Bernhard, M.: Efficient Methods to Account for Cloud-Top Inclination and Cloud Overlap in Synthetic Visible Satellite Images, J. Atmos. Ocean. Tech., 35, 665-685, doi:10.1175/JTECH-D-17-0057.1, 2018].

[Reply]
We agree with the fact that the reflectance bias was made between the observed reflectance of "visible" satellite instrument and the modelled reflectance from the numerical model output in Scheck et al. (2018). We added this paper in the reference list as follows:
[Old, lines 221-223]
"In general, the biases are estimated using the time averaged departures between the observed radiances and the simulated radiances from the spatiotemporally collocated model field (background), because of

the absence of reference data suitable to compare the satellite observations."
[New]
"In these experiments, the biases are estimated using the time averaged departures between the observed radiances and the simulated radiances from the spatiotemporally collocated model field (background), because of the absence of reference data suitable to compare the satellite observations (Scheck et al., 2018)."

Scheck, L., Weissmann, M., and Bernhard, M.: Efficient Methods to Account for Cloud-Top Inclination and Cloud Overlap in Synthetic Visible Satellite Images, J. Atmos. Ocean. Tech., 35, 665-685, doi:10.1175/JTECH-D-17-0057.1, 2018.

2) In my understanding, the bias of the observation could be estimated by this method, i.e, averaging the departures between the observed radiances and simulated radiances, only when the simulated radiances are unbiased statically. If the simulated radiances are biased, they cannot represent the "truth" very well. The simulated radiances are strongly influences by the model output of the pre-trial run. As is mentioned in L422-428, the model output of the pre-trial run is biased. Therefore, it seems that the estimated bias contains the observation bias and some model bias. About this problem, could you please give some explanations here?

[Reply]
I agree with your opinion. The bias correction method applied in this study basically uses the departures between the observed radiances and the simulated radiances from the spatiotemporally collocated model forecast field (i.e., 6-h forecast). As you mentioned, for this correction method to work correctly, the simulated radiances should be unbiased, meaning that the model field (6-h forecast) is also bias-free. However, as it is still uncertain if the model fields are unbiased, many efforts have been taken to use the well-calibrated in-situ observations as the reference data. Nevertheless, it is not easy to find the reference observations able to compare the satellite radiances. For example, it seems to be the best way to compute the bias between the satellite-observed microwave radiances and the simulated radiances from the collocated radiosonde measurements (i.e., temperature and water vapor profiles) using the radiative transfer model. In this case, the number of radiosonde measurements is not enough to compute statistically significant values for the short-term period. In addition, the simulated radiances could be biased because the radiative transfer model (RTM) still has uncertainty, which is used to simulate the radiances from the radiosonde measurements. The research community related to satellite data assimilation has attempted to separate the integrated bias into the instrument bias and the model bias, but this bias correction issue is still not solved. Currently, thus, the biases are practically estimated using the departure between the satellite observations and the simulated observations from the model fields. To describe this point, we revised some sentences as follow:
[Old, lines 221-224]
"In general, the biases are estimated using the time averaged departures between the observed radiances and the simulated radiances from the spatiotemporally collocated model field (background), because of the absence of reference data suitable to compare the satellite observations."
[New]
"In these experiments, the biases are estimated using the time averaged departures between the observed radiances and the simulated radiances from the spatiotemporally collocated model field (background), because of the absence of reference data suitable to compare the satellite observations (Scheck et al., 2018). The use of the simulated radiances from the model background (i.e., 6-h forecast) may be questionable because the model background could be biased. However, it is effectively impossible to

find sufficient reference observations for comparing with these satellite observations, so the biases are made using the departures between the observed radiances and the model simulated radiances.”

3. L244: I was confused about the formula (4). Why was the averaged residual scan bias obtained by removing the mean bias of two near nadir FOVs (15 and 16) from the bias of the departure for each FOV (1–30)? In other words, why the off-nadir bias was estimated by subtracting the near-nadir bias? Could you please give more details here?

[Reply]
Thank you for your comment. As mentioned in the manuscript, the AMSU-A instrument is a cross-track microwave sounder that scans 30 FOVs per horizonal scan line with the scan angle between ±48.33°. Thus, the optical path length between the earth and the instrument varies depending on the scan angle, called the limb effect. Considering that the scan bias correction aims to only remove the biases caused by the incorrect limb effect in the radiative transfer modeling, the bias in the near nadir (15 and 16 scan position) is mainly due to the air-mass bias, not the scan bias because the scan angles are almost zero. For this reason, as shown in Eq. 4, the residual scan bias is computed by extracting the averaged bias of two near-nadir FOVs (15 and 16) from the bias of each FOV (1–30).
To make this point clear, we modified some sentences as follows:
[Old, lines 238-239]
“Second, the averaged residual scan bias is obtained by removing the mean bias of two near-nadir FOVs (15 and 16) from the bias of the departure for each FOV (1–30).”
[New]
“Second, as the scan bias derived from the departures between the observed radiances and forward-modeled radiances likely includes the air-mass bias, the averaged residual scan bias is obtained by removing the mean bias of two near-nadir FOVs (15 and 16) from the bias for each FOV (1–30).”

4*. L337: Are twenty ensemble members enough for data assimilation in a global model? Were all ensemble members configured with the same physics options but with different initial and boundary conditions?*

[Reply]
All ensemble members have the same model configurations (e.g., model dynamics and physics), but are initialized from different initial/boundary conditions (i.e., atmosphere, land, and sea ice components). However, for the ocean component, the same sea surface temperature (SST) dataset is used for all ensemble members.
And, as you mentioned, the assimilation techniques based on the ensemble Kalman filter (EnKF) cannot be free from the sampling error that is caused by the limited size of the ensemble members. It is a fact that the larger the number of ensemble members, the higher the statistical significance of the correlation derived from the ensemble spread. However, as the number of ensemble members strongly depends on the computation capacity available to run the ensemble-based numerical model, it is difficult to increase the number of ensemble members indefinitely, in particular, in the small research community in which the high-performance computing system (e.g., super computer) is not installed. Thus, to eliminate the unreliable correlation derived by insufficient ensemble members, the localization method is applied. In addition to the localization, the sampling error correction scheme is available in DART, which uses pre-defined information about the correlation between the model state variables and the observations as a function of ensemble size. And, there are some previous studies in which 20 ensemble members were applied in the EnKF technique to assimilate the satellite-based output. In Mizzi et al. (2016), the

retrieval data of the satellite-based atmospheric composition were assimilated using the EnKF technique with twenty ensemble members. In addition, Zhang et al. (2021) assimilated the carbon dioxide (CO_2) retrievals from the Orbiting Carbon Observatory 2 (OCO-2) satellite into DART where the size of ensemble member was set to be twenty.

Mizzi, A. P., Arellano Jr., A. F., Edwards, D. P., Anderson, J. L., and Pfister, G. G.: Assimilating compact phase space retrievals of atmospheric composition with WRF-Chem/DART: a regional chemical transport/ensemble Kalman filter data assimilation system, Geosci. Model Dev., 9, 965–978, https://doi.org/10.5194/gmd-9-965-2016, 2016.

Zhang, Q. W., Li, M. Q., Wei, C., Mizzi, A. P., Huang, Y. J., and Gu, Q. R.: Assimilation of OCO-2 retrievals with WRF-Chem/DART: A case study for the Midwestern United States, Atmos. 985 Environ., 246, 118106, https://doi.org/10.1016/j.atmosenv.2020.118106, 2021.

*5. L344: A half-width of 0.075 radians is equivalent of a localization distance of 955.65 km (2*6371*0.075) in the horizontal direction. Could you please explain why is a localization distance of 955.65 km was set?*

[Reply]
Similar to the determination of AMSU-A spatial thinning distance (in 1st comment), a localization half-width of 0.075 radiance in this study was also chosen through three trial experiments. In three experiments, three localization half-widths of 0.15 (a default value), 0.075, and 0.0375 were applied. Among them, the largest analysis improvement was shown when the localization half-with was set as 0.075. A localization distance of about 1000 km (0.075 radians) also corresponds to a typical horizontal length scale of atmospheric synoptic phenomena (e.g., cyclones). To make this point clear, we revised some sentences as follow:
[Old, lines 344-345]
“In this study, the horizontal/vertical localization half-width of 0.075 radians was employed to prevent the use of erroneous correlation.”
[New]
To determine the localization half-width, three extra assimilation experiments were run with different half-widths (i.e., 0.15, 0.075, and 0.0375). As the largest analysis impact was made with the half-width of 0.075, the horizontal/vertical localization half-width of 0.075 radians was employed to prevent the use of erroneous correlation.”

6. Figure 6 and Figure 11: Since different variables could interact with each other, as is mentioned in L383, I think it would be interesting to see how the humidity-related variables were influenced by assimilating the AMSU-A observations.

[Reply]
Thank you for your comment. As the number of radiosonde relative humidity is quite small in the conventional observations obtained from the NCEP ADP, it is difficult to diagnose the model first-guess departure of the humidity against the radiosonde humidity measurements. Instead, we compute the error of specific humidity (g/kg or kg/kg) in the analysis against the ERA5 reanalysis to assess the impact of assimilating the AMSU-A observations on the model humidity field. As shown in Fig. S1b, the positive analysis impact (about -5 %) is shown in the troposphere (1000 hPa – 200 hPa) over the Northern Hemisphere where the analysis impact on the primary variables (i.e., 500 hPa geopotential height, temperature, zonal wind, and meridional wind) is significant. In the tropics and Southern Hemisphere,

the analysis impact is slightly negative, but not statistically significant, except in the lower stratosphere (near 100 hPa) in which the large negative impact (about 18 % and 15 % in the tropics and Southern Hemisphere, respectively) is shown. However, as shown in Fig. S1a, the magnitude of the standard deviation, a denominator to compute the normalized difference of the standard deviation (Fig. S1b), is quite small in the lower stratosphere for the control run. Thus, the change of the standard deviation seems to be negligible in the lower stratosphere. We mention this point in the revised manuscript as follows:

[New]

"In the model humidity field, a positive analysis impact only occurs in the Northern Hemisphere (not shown), but is not as significant as the abovementioned parameters (i.e., 500 hPa geopotential height, temperature, and winds). As a further study, we plan to assimilate the Microwave Humidity Sounder (MHS) providing information on the vertical structure of humidity so that the initial condition of model humidity is improved."

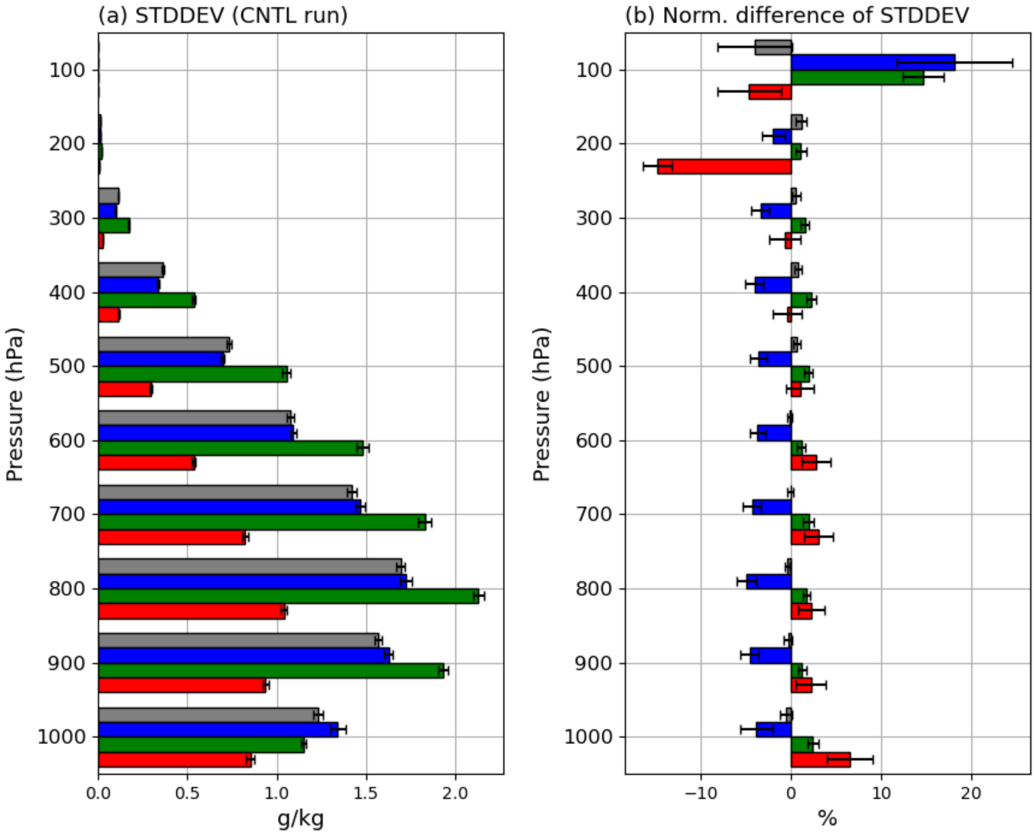


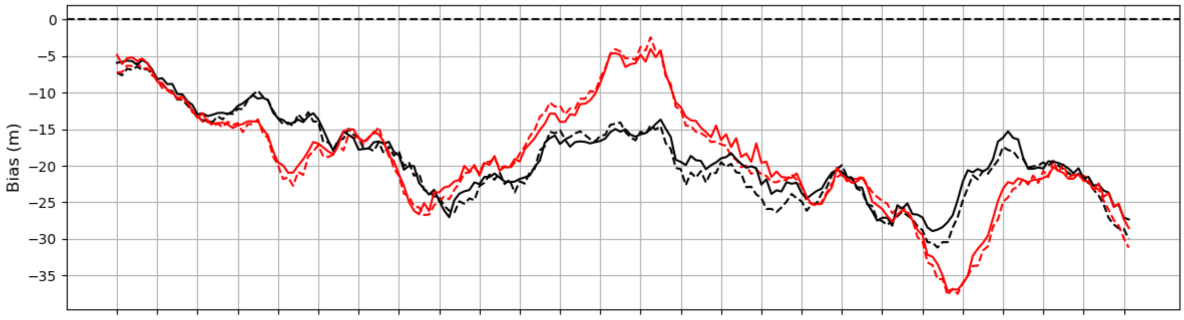
Figure S1. (a) The standard deviation of specific humidity (g/kg) for the control (CNTL) run and (b) normalized difference of the standard deviation of specific humidity between the experiment (AMSU-A) run and the control (CNTL). The 99% confidence intervals are indicated by the horizontal black lines.

7. The assimilation would generate positive impact on the model variables at the analysis time. After that, the positive impact could diminish quickly with model integration since the balance law between different variables was not respected during the data assimilation process. It would be really helpful to see how the errors (biases, STDDEVs) for both the analysis and forecasting fields vary with time.

[Reply]

In general, the forecasts are generated from the analysis field derived at each analysis cycle in the data assimilation system comprising the forecast system and the data assimilation system. In the current version of the CESM/DART data assimilation system, long-range forecasts are not generated yet. However, the short-range forecast (i.e., 6-h forecast) is used as the background at each analysis cycle in the DART assimilation system, which is archived. As shown in Fig. S2a, the global-mean bias largely fluctuates during the trial period (11 August – 30 September 2014), but is negative for two runs overall. The STDDEV values fluctuate between 30 m and 70 m, but the AMSU-A run has an overall small STDDEV compared with the CNTL run, except for the period from 1 September to 9 September 2014 (Fig. S2b). As the DART reanalysis was used as the starting initial condition on 0000UTC 11 August 2014 for both runs, the STDDEV rapidly increases for the spin-up period (0000UTC 11 August to 1800UTC 24 August 2014). This pattern is also shown for the short-term forecast (6-h forecast). After the spin-up period (two weeks), all model variables are adjusted to be more consistent with the observations, which presumably represent a nearly balanced atmosphere, so at least some of the balance is preserved. In addition, as shown in Fig. S2, the analysis increments, which are the changes to model variables caused by the assimilation, are small, so any imbalance will be small.

(a) 500hPa GPH Bias



(b) 500hPa GPH STDDEV

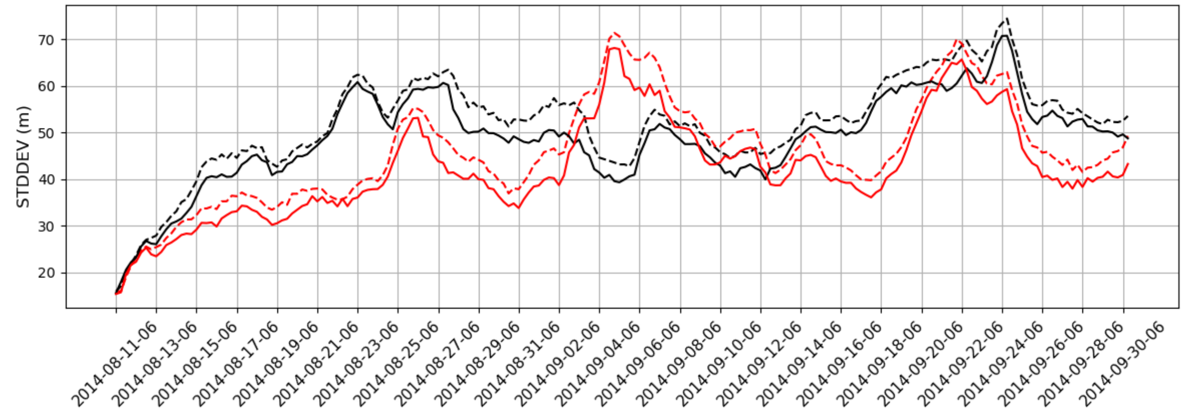


Figure S2. Time series of (a) the bias and (b) standard deviation of 500 hPa geopotential height for the control (CNTL; black) run and experiment (AMSU-A; red) run during the trial period from 11 August to 30 September 2014). The solid and dashed lines indicate the errors (i.e., bias and standard deviation) of analysis and 6-h forecast (background), respectively.