Review comments to gmd-2015-203**:**

This paper presented the application of Multi-Model Ensembles (MME) for the thermospheric density forcasting. First, the neutral density from three individual models has been compared with CHAMP satellite observations. The MME has been constructed from three models, which shows a significant improvement compared with individual models. TIE-GCM has also been run for forecasting and initialized with MME every 6 hours The results are very interesting and relevant to the current topics. The paper is suitable for the publication of GMD. However, some methodology is not very clear and the clarification is needed.

(1) Line 202: "The model all perform very similarly": To my eyes, the model results in Figure 5 are quite different, especially the variation of neutral density during the storm period, even the standard deviation and correlation are similar. Does it indicate the limitation of using standard deviation and correlation to judge the performance of simulation?

(2) Line 205-209 and Line 251: the performance difference between the third scenario and the first two is mainly explained as the consequence of the solar activity variation. However, F10.7 actually represents the solar irradiation worse during last extremely quiet solar minimum in 2008 and 2009 than the solar maximum in 2001. It is not clear that the comparison is necessarily worse in the solar maximum when the models are driven by F10.7. Meanwhile, the third scenario includes a much larger geomagnetic storm than the other two, which may also contribute the performance difference in addition to the solar activity change.

(3) Line 296: "Using the MME as the initial condition in TIE-GCM …": The terminology of "initial condition" is confusing. Typically, the initial condition is a one-time thing for the simulation, which is used at the beginning of the simulation period. What has actually been done in this study is to retune TIE-GCM to MME every 6 hours, which is probably different from the initial conditions people usually talk about.

 (4) Table 2: The weights are quite different from one scenario to another. It may indicate that the weight MME may not be applicable for forecasting using the weight calculated from historic events.

(5) The abstract needs to represent the content of the paper better by including more information about the approach and main conclusion.

(6) Figure 20: the label shows the period of Nov. 2008, which is the second scenario.

(7) Line 252: t → It