Comment made by: Anonymous Referee #1
Received and published: 28 April 2020

Reply by authors is shown in blue and start with the symbol >>.

General comments:
This manuscript is very well-written and presents the results in a straight-forward manner. It is clear that assimilating various retrievals improves model representation of the three instability variables described. I only have a couple comments outlined below regarding the methodology details.

>> Thank you for these supportive words.

Comments:
Section 2.3.2: This section is very brief on the details of the bootstrapping technique. Perhaps include more details and some references?

>> The bootstrap operates by constructing the artificial data batches using sampling with replacement from the original data. Conceptually, the sampling process is equivalent to writing each of the N data values on separate slips of paper and putting all N slips of paper in a hat. To construct one bootstrap sample, N slips of paper are drawn from the hat and their data values recorded, but each slip is put back in the hat and mixed (this is the meaning of "with replacement") before the next slip is drawn. This process is repeated a large number of times yielding, for example, to 1000 samples of size N that are slightly perturbed versions of the original data set. The 95% confidence intervals for the different statistical indicators can then be derived from their P975 and P025 observed percentiles of their distribution as obtained from the 1000 perturbed series. More mathematical details can be found in Wilks (2011), Efron & Gong (1983) and Downton & Katz (1993).

>> Coming back to our manuscript, the bootstrap technique was applied to the temporal analysis of each index. In our case, the original time series used in the Taylor diagrams consist of 60 values, each of them for the corresponding month along the period 2010-2014 (12 months x 5 years). For the bootstrap, we created 1000 perturbed time series taking into account different samples of the data. 67% of the new time series (2/3 of the length of the original time series - 40 values in our case) is made from the original data, and the remaining 33% (1/3 - 20 values) is chosen from those values already taken from the original data. For each correlation calculated, the same samples are taken from the observed and model data.

>> In order to clarify how the bootstrap is performed in our analysis, some extra lines will be added to section 2.3.2 (Analysis) of the paper. This new lines will sum up the information presented above, and they will include the citations.

>> Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

>> Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37(1), 36-48.
DOI: 10.1080/00031305.1983.10483087

>> Downton, M. W., & Katz, R. W. (1993). A test for inhomogeneous variance in time-averaged temperature data. Journal of climate, 6(12), 2448-2464.
DOI: 10.1175/1520-0442(1993)006<2448:ATFIVI>2.0.CO;2

What is exactly being verified? Model analyses after assimilation cycling or forecasts? This is not abundantly clear. If these are forecasts being verified, how do the statistics vary with lead time?

>> The structure of segments used in each experiment is presented in Figure 1. The N experiment is generated running 6-hour long segments that are restarted from the restart file produced at the end of previous segment (top panel of Figure 1). This is similar to a continuous WRF run where the boundary conditions (in our case, from ERA-Interim) are provided to the model every 6-hours after the initialization of the model the 1st of January, 2009.

>> For the experiment including the data assimilation, the structure is a little bit more complex. In these case, 12-hour long segments starting at every analysis time (00, 06, 12 and 18 UTC) are used (bottom panel of Figure 1). The analysis are generated from the outputs of the model at a 6-hour forecast step from the previous segment as first guess in a 3DVAR data assimilation scheme. The data assimilation is performed using the observations in PREPBUFR format obtained from the NCEP ADP Global Upper Air and Surface Weather Observations (ds337:0) dataset generated by NOAA. Only those observations included in a 2-hour time-window centered at the analysis times were included.

>> In both cases, the outputs are saved every 3 hours, which means that analysis (00, 06, 12 and 18 UTC) and 3-hour forecasts (at 03,09, 15 and 21 UTC) are included in our results. These recording frequency is highlighted with magenta ellipses in Figure 1. These are the data that are verified in the manuscript.

>> The new version of the manuscript will include an expanded explanation about how both simulations were created in order to make it clearer to the readers.

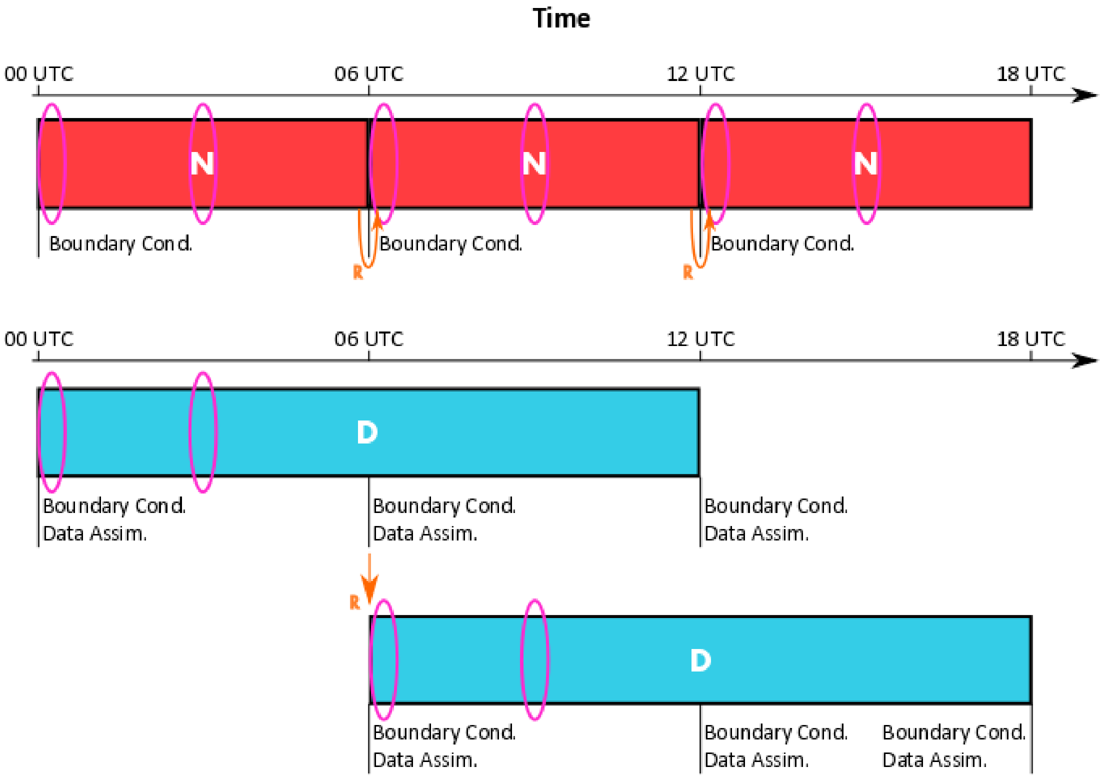

Figure 1: Diagram showing the structure of the segments used for running N (top, in red) and D (bottom, in blue) experiments with the WRF model. The outputs recorded are highlighted with magenta ellipses, and the restart files used to run the segments are shown in orange.