General comments

This study aims to evaluate different techniques involving ingestion of vegetation, soil temperature and moisture, and SST data from satellite and LDAS output, to improve high-resolution WRF simulations for use with the CMAQ air quality model applied to the LMOS field study. The study also includes two different physics configurations that alter the LSM, PBL, and microphysics. Unfortunately, the model configuration designated "EPA" seems flawed relative to our experiences modeling many fine scale domains. How it may be flawed, however, can only be speculated. I think more should be done before this "EPA" simulation is used. We understand that many users may not fully appreciate some of the details in configuring inputs and some namelist settings (i.e., Obsgrid) that need some scale awareness, but we think that a sensitivity study of different physics configurations should make sure the model has the most appropriate inputs and runs correctly. Instead, the paper seems to suggest that these results are just what you get when running the "EPA" configuration. Also, calling what they have done the "EPA" configuration makes it seem like they were either performed by the US EPA or endorsed. The EPA has used similar configurations at high-resolution with good results as shown in the recently published LISTOS modeling study (Torres-Vazquez et al. 2022). The LISTOS area is just as complex as LMOS with land-sea breeze impacts. The LISTOS modeling showed that the 2-m temperature RMSE for the 1.33 km domain were clearly lower than the 12 km (1.75 K for 12 km and 1.60 K for 1.33 for the whole summer) which is also much lower than the LMOS results shown here. The reasons for the poor results may be related to the data assimilation procedures/inputs used in this study. However, it is not clear from the manuscript what was done in some key areas of data preparation.

It is mentioned that the indirect soil moisture and temperature nudging was used in the PX LSM. Some details are not clear. It seems that they used the 0.25-degree resolution GFS Final re-analyses available at 6-h intervals as background to re-analyze with MADIS observations using OBSGRID. However, these analyses are much too coarse spatially and temporally to give good results for WRF runs at 1.33 km resolution which would explain why the results are not as good as the runs used for LISTOS which used the UnRestricted Mesoscale Analysis (URMA) data set, available in as fine as 1-hourly increments on a 2.5 km grid spacing for 1.33 km WRF runs. Some of these ideas for best practices for running the PX LSM at high resolution are described in a document http://www2.mmm.ucar.edu/wrf/users/docs/PX-ACM.pdf referenced in the WRF users' guide https://github.com/wrf-model/Users_Guide. However, this alone is unlikely to have resulted in the large errors and biases especially in humidity that are shown in the paper. Therefore, I think there may have been other issues like potentially problems with the observation nudging.

It is stated several places that surface observation nudging was used for the "EPA" runs, but the details of how this was done are not described. On lines 470-472: *The superior performance of the EPA simulation on the 12-km domain is partially an artifact of its use of surface observation nudging because the same observations used in the nudging routine were also used for verification.* If the same observation nudging was used for the 1.33 km runs the results would also show superior performance at the observation locations. Seems something went wrong and likely an issue with the nudging itself. Until the flaws in the "EPA" runs are investigated, discovered, and corrected, I suggest that these simulations be removed from the study. It seems like the experimental design implicitly linked obs nudging to the US EPA configuration. We think this should have been a separate sensitivity.

The comparisons of the various YNT sensitivity simulations show relatively small differences. There are some interesting analyses of the differences and discussion relating the differences to the variations in the use of soil, vegetation, and SST data sets and a change in the height of analysis FDDA. However, these discussions repeatedly refer to comparisons to the flawed "EPA" runs.

Specific comments:

Line 20: Please do not refer to the simulation using PX and ACM2 as "EPA" since the EPA was not involved with these runs.

Line 24: "physically unrealistic ground heat flux". This should have been a clue that something was seriously wrong and should be investigated.

Line 93: This is a 7 year old version of WRF. Why not use a newer version? There have been significant changes and error corrections in PX and ACM2 since then. Noah has also been updated.

Line 111: misleading. Better to say local and non-local vertical transport

116: "available surface observations" is not 100% accurate here. The indirect nudging uses the surface analysis in the wrfsfdda_d0* files. Obsgrid creates the wrfsfdda_d01 file using the GFS-FNL here + MADIS point obs, but it is more accurate to say the indirect soil nudging uses a surface analysis from Obsgrid.

121-123: More explanation is needed especially about the background analysis

144: If I understood correctly some of these datasets are use daily not just for initialization

Table 1. Please clarify what is meant by "obs nudging to MADIS". Is this direct obs nudging or the indirect soil T and SM nudging using obsgrid?

231-233: Again, need to explain surface nudging

270-272: Basic rule of thumb: don't use coarser analyses to nudge finer models

283-284: This is unreasonable performance. The simulation should be fixed or removed from comparison.

310-311: This suggests deficiencies in YNT physics

334: The fact that WS errors were smaller for "EPA" than T2m and q-2m suggests that the data used for soil nudging was very detrimental.

362-365: This shows some serious errors in the "EPA" runs. Need to figure that out or remove from comparisons.

405: This assumed surface energy balance neglects Cp dTg/dt. Why not sum the up and down SW and LW components? That would be net surface radiation.

425-434: These results seem to indicate that there is a larger problem than the low resolution of the nudging data. The Gflux always < 0 and the large wet bias day and night suggests that the nudging data is much too cold. If a DA scheme uses bad data, the results will be bad. The surface nudging analyses should be evaluated the same way as the model output. I expect this will show large errors in the nudging data. This might suggest errors in the OBSGRID processing.

445: This is contrary to what the LISTOS study showed

460: Again, please clarify “surface observation nudging”

462: Noah is not more sophisticated than PX LSM overall. While Noah has more soil layers, PX uses more refined representation of land use related parameters. Also, the indirect soil nudging capability is a big advantage when applied correctly as has been demonstrated many times in many publications. In other ways the two models are roughly equivalent.

464: Again, not just initialization.

470-472: if the good performance of the 12 km “EPA” run is due to obs nudging then why did the 1.33 km run not have similarly good results. Observation nudging using the same obs as used in the statistics should give much, much better results.

474-476: I agree that the relative spatial sparseness of the obs do not accurately capture and constrain small-scale features on the 1.33 km grid but evaluating wth the same observations as used on the obs nudging should result is very close agreement regardless of the model grid scale. This suggest some potential problem with the obs nudging that should be investigated, or removing Obs nudging from the experiment or as an independent sensitivity

485-488: The two-layer soil structure is designed to work with the indirect nudging. This normally works well and gives better results that other LSMs when applied correctly.