This is a well-written study intercomparing 4 ET algorithms against FLUXNET ET measurements. Like the rest of the group of papers coming out of this team (GEWEX/LandFlux, WACMOS), the strengths are in the selection of algorithms, the common forcings, and the rigorous analyses. Similarly, the weaknesses include the fact that the results are scattered among different papers with somewhat different details of analyses, so it is very difficult to understand the cohesive picture, and that the papers do little to go beyond statistical intercomparison and into the realm of science understanding.

Detailed comments:

- Nomenclature consistency: Mu et al., 2011 abbreviation is referred to inconsistently across projects, i.e., PM-Mu, PM-MOD, PM-MOD16, etc. Same goes with GLEAM (colon/no-colon; Methodology vs. Model).
- It should be made clear how this study advances past Vinukollu et al 2011.
    - Also, please make clear how this is *scientifically* different than the Michel paper (in prep at the time of this review writing, but soon to be in Discussions). At first, when reading the Michel paper, I thought the main difference was the 3-hourly analysis, but then I've seen the McCabe paper also includes 3-hourly…
    - Speaking of which, given the whirlwind of papers coming out of this GEWEX/LandFlux & WACMOS group (e.g., Michel, Miralles, McCabe, Ershadi, …), I strongly urge McCabe in particular to write a meta-analysis/review paper of these papers to distill everything down into 1 place (include the Vinukollu, Jimenez, Mueller, etc. papers too). Aim high (e.g., one of the Natures, etc., or perhaps WRR).
- A semantic nuance that would improve the interpretation of the results further would be to rephrase/reframe model performance not so much in that X model overestimates/underestimates, but that it's actually the model in conjunction with the selected forcings. E.g., it may not be inherent to the model itself that it is biased high or low, but rather due to the forcings. This would primarily be for bias, not as much for the other statistics, though the other statistics would not necessarily be completely immune either.
- How can error be reduced in the models further? What causes the error? I think a lot of the error that the authors attribute, as calculated, to the models is in fact error in the data. It remains an outstanding question in this analysis why a model would do well at some sites, but not well at other very similar sites. Or, even inconsistently throughout time within a single site.