Review: Evaluating the E3SM Land Model Version 0 (ELMv0) at a temperature forest site using flux and soil water measurements

Authors: Liang, Wang, Ricciuto, Gu, Hanson, Wood, Mayes

Synopsis: The authors run an out-of-the-box simulation of E3SM (which is the same as CLM4.5-CN; I'm not sure what makes E3SM distinct) and compare results to Eddy Covariance (EC) fluxes and soil moisture observations. The default model is found to have Gross Primary Productivity (GPP) and Soil Respiration (SR) that are too low when compared to observations. Near-surface Soil Water Potential (SWP), calculated using relationships that determine SWP from Volumetric Water Content (VWC) based on Clapp&Hornberger result in potentials that are too high during the winter, and too low in the summer. Overall, SWP is too low at low moisture in the near-surface soil, and was slightly high at depth when moisture content was higher.

Five different treatments for relating VWC-SWP were tested, and the model of Hanson was found to have the smallest errors when compared to observations. However, when C&H was replaced with Hanson GPP was slightly low and SR slightly high when compared to observations, and the model did not reproduce either the amplitude or sign of interannual variability. Therefore, coefficients influencing Specific Leaf Area (SLA), fractional leaf N used in Rubisco and several coefficients controlling leaf senescence were changed, and results were improved in evaluation of mean seasonal cycles of LAI, GPP and SR.

Finally, there was speculation about which mechanisms and processes might be responsible for model-data mismatch after the aforementioned tuning was complete. These include model Q10 for heterotrophic respiration, microbial biomass and seasonality, and macroinvertebrate (earthworm) influence on carbon cycle processes, and root exudates. The authors exhort the community to pay particular attention to SWP in simulations, and to consider inclusion of these added processes in models.

Review: One could consider this a model-tuning paper. A default model was run, deficiencies were noted, and changes were made to parameters and model physics. This is fine, and has been done many times previously (e.g. Sellers et al. 1989), but I'm not sure that the present paper really tells me anything about how the world works. I work with models that simulate land-atmosphere interaction, and there is nothing in this paper that makes me want to look at my model code and start performing tests and making changes. Hanson worked better, but it worked better at *one place on the planet,* at a particular deciduous forest (DBF) in the North American midwest. I suspect that if we were to perform evaluations like this at multiple EC sites (across multiple DBF sites and across multiple PFTs), I expect that we would find that each of the VWC-SWP treatments would come out on top at least one or more times. We would also likely find that the SLA, Nitrogen and senescence parameters could take multiple values as well.

My main complaint now levied, I will also say that just because this paper does not particularly *excite* me, there is nothing *wrong* with the analysis. The paper follows a logical progression, and the analysis and presentation of results is done professionally and is easy to follow. I think there is value in the paper, and my official recommendation is to accept this manuscript with minor revisions.

The paper is quite short (14 pages), which is nice, but I think there might be some expansion of analysis and explanation that would add value to the research.

Merely stating that the model was unable to capture observed response to the 2012 drought is extremely unsatisfying. This is an *opportunity* to explore model behavior, and perhaps gain valuable insight into processes and mechanisms. I find it interesting that observed near-surface (10cm) SWP (Fig 3b) was not exceptionally low in 2012; the year did not look much different from 2011 or 2013, and in fact looked wetter than 2005-2007. That is interesting; what was deep soil SWP doing in those years? From Figure 2, we see that observed SR oscillated up and down between 2005-2007, while GPP dropped from 2005 to 2007. What was the model doing? What did BTRAN look like in 2012, as compared to other years? How about LAI? Is there a near-surface water table in the simulations that prevents root stress? Are there constraints on stomatal conductance due to high VPD or unfavorable temperature? What did they do in 2012? In fact, simulated SR and GPP both increased from 2011 to 2012, while there were dramatic drops in the observations of both. If the model does not respond to the drought, you should be able to tell your reader *why*, and speculate whether that behavior is realistic or not, and how that behavior might impact model performance in other years. I would like to see some exploration of IAV, and explanation of why Hanson provides an upgrade from C&H in this regard.

Related to the above is the fact that in 2007 there was a significant drop in observed SR and GPP when compared to 2006. The default model (C&H) showed drops that are more similar to the amplitude of the observed reduction, even if there is an offset or bias. In fact, the Hanson model shows almost no interannual variability (IAV) in SR and GPP at all. Is this really an improvement? One might make the case that you would have a better simulation of the observed flux by increasing Vcmax (and perhaps SLA and the senescence parameters) in CLM without changing the soil; your GPP would go up, which would translate into larger carbon pools and subsequent increased RS. You would also retain a more realistic comparison with observed IAV. Is there a reason to suspect that this would not work?

In section 4.1 two paragraphs (lines 30-33 on page 11, continued in lines 1-7 on page 12; lines 8-13 on page 12) describe how C&H was developed from textural classes and not sand/clay fractions, and how models might make use of near-continuous SWP observations. I'm not sure what these 2 paragraphs bring to the analysis, since neither has been done. Do they merit this much attention?

The authors state that "SWP in simulations in ESMs should be calibrated carefully with observations…", but this is clearly impossible and unrealistic in global simulations. If the ultimate goal is correct simulations of global biogeophysical behavior, then we have a disconnect between what the authors are doing here (tuning at a single site) and what we are told is the ultimate goal (accurate representation of global carbon cycle). This is a persistent and real problem. We calibrate our models on site-level data and then extend that behavior to the globe. I'd be interested in some discussion of how we might use site-level studies to improve global simulations.

Specific Comments:

- Equation 4: The subscript should be liveCroot, should it not?
- Page 4, line 9: coarse
- Page 5: is 'residual' water content the same as wilt point?
- Table 1: I'm not sure what AIC is: shouldn't it be explained, even if briefly?
- Page 8, line3: I know what 'btran' is, but some of your readers may not. You should explain this variable.
- Figure 4, Figures S6-S7: Make lines darker, shading lighter. Hard to discern individual simulations.
- I might have missed this, but what is the porosity and sand/clay content of the soil at the MOFLUX site? If VWC at depth regularly drops to between 15-20% (Fig 1b) then it must have considerable sand content. My recollection of more clayey soils is that wilt point will be much higher. Is this soil representative of the region and/or PFT?
- Equation 9: It appears that the environmental modifier for water has value of 0 and low water (conditions too dry for microbial activity), varies between 0-1 for moisture up to (PSI)max. What is the difference between (PSI)max and (PSI)s? Are they the same? Most models I am familiar with will have an 'optimum' soil water content or potential for respiration, the idea being that either too dry or too wet (anaerobic) conditions are unfavorable for microbial decomposition of carbon stocks. The 'too wet' does not seem to be the case here. Why is that?
- Using 10 years of tower forcing to perform a 200-year spinup of carbon pools concerns me. I understand that this might be all the tower data available, but, especially for carbon pools, I'm concerned that anomalies in the 10-year meteorology may be aliased onto pool size. Did you consider using a reanalysis product (CRU, NCEP, ECMWF) for spinup and then use the tower data for the transient run?

References:

Sellers, P.J., W.J. Shuttleworth, J.L. Dorman, A. Dalcher, J.M. Roberts, 1989: Calibrating the Simple Biosphere Model for Amazonian Tropical For- est Using Field and Remote Sensing Data. Part I: Average Calibration with Field Data. Journal of Applied Meteorology, 28, 727-759, August

1989.