The manuscript by Byrne et al is an analysis of seasonal cycles of carbon fluxes over northern ecosystems, with a comparison of inversion results of NEE against DGVM estimates of NEE, and then a comparison of the inversion NEE estimates minus GPP and an estimate of Ra to infer Rh, against DGVM estimates of that as well. The main findings are (a) that NEE late-season positive fluxes are higher in the observations than the models, and (b) that the inferred Rh seasonality indicates that the DGVMs underestimate that late-season Rh.

Overall, I think the comparison between inversion results and models is really useful, and the paper should be published. But I find it an interesting but not entirely satisfying analysis. One problem is that the number of different steps from NEE to Rh seems like it introduces the potential for several errors to creep in, particularly as relate to Ra. Second, there are any number of reasons why the DGVMs could show a bias relative to the observations, and it is certainly possible that the lack of deep-soil respiration is one reason. But the attempt to provide a mechanistic explanation here using a simple model is not very clear, and subject to somewhat arbitrary choices like how to handle substrate seasonality. I wonder if a slightly different approach of looking at the DGVMs themselves, and asking whether there are structural or parametric characteristics of the models that govern the shapes of their seasonal cycles, and which might provide some clues for identifying whether any of them do a better or worse job than others?

Line 12: please provide uncertainty range for the DGVM estimates, as you do for the data-driven estimates

Line 17: "is not well captured by current DGVMs." Any DGVMs, or just the ensemble mean?
Line 70-72: Could you clarify whether you are using monthly mean CUE values or annual mean values here?

Lines 99-100. How confident are we in the soil temperature predictions of these models? There have been a few analyses of the soil temperature dynamics and permafrost statistics of climate models at high latitudes. Does this set represent a set of best-performing soil temperature models?

Fig. 2. I think that the per-area fluxes are more meaningful here, otherwise the reader gets the suggestion that NEE is higher during the summer in the colder than the warmer regions, which is confusing. So I suggest switching fig. S7 and fig. 2, and in general reporting things per unit area.

Figs 2 & S7: I am skeptical about the errors introduced by the GPP -> NPP conversion, I think it would be useful to include a set of GPP panels as well, since, like NEE, that is the most directly observed, with the NPP and RH much less direct.

Figs 2 and S7: A lot of the focus of the discussion is on the autumn differences, but I wonder if the more general problem is that the winter respiration in general is underestimated by the models in the cold region. This would be consistent with the findings of Natali et al., but given the larger-scale datasets used here would still be an important point to emphasize here.

Lne 264: This isn't really a shift, so much as a bias in TRENDY relative to the observations?

Lines 164-270 and fig. S12. I think FLUXNET is actually telling a different story than the larger-scale datasets. The TRENDY models actually have a higher positive NEE anomaly during the shoulder season than FLUXNET, which is the opposite pattern shown in fig. 2c. If this is correct, then I think the discussion of this result needs to be revised accordingly.

Section 3.2, I'm not sure I understand what new information the 14-day-resolved data provides beyond what is in the monthly data. Is this analysis really necessary? If so, could the authors give a bit better motivation and explanation?

Fig.3. I'm very skeptical about how narrow the range of uncertainty in panels a-c are here. What is that a measure of?
Lines 334-348, and figure 4. I don't understand this sensitivity analysis, or why the seasonal cycles in panels g-i are so different from the ones in panels d-f. Could you clarify a bit more what is being shown here? Further, the argument about deep soil playing a greater role should help with the autumn respiration peak, but less so with the bias in respiration in the cold region throughout the winter. What does this analysis have to say about that?

Line 441. I don't understand the line "TRENDY v8 data were downloaded from trendy-v8@trendy.ex.ac.uk.", since that is an email address, not a URL. Please provide a URL or DOI to a FAIR-aligned data archive where the data can be freely downloaded or, if the data is not available, then per this journal's data policy, a detailed explanation of why this is the case is required.