I thank the reviewers for thoroughly addressing my concerns. I think the paper is now basically ready for publication. I have two comments though, which I want to give the authors a chance to address, so am recommending Minor Revisions.

Best wishes,
Kristian Strommen

MINOR COMMENTS

1. Section 2.3: Thanks for taking the time to explain this more. What I now understand to be the key point is that the difference of 1 day in the initialisation dates can actually have bigger impact on your estimates than I initially guessed, due to the speed of propagating transients. It therefore does make sense that rather comparing biases against ERA5 may be more helpful, since you remove a lot of the variance contributions from these changing transients. I would however like to make a suggestion regarding this Section.

   Firstly, edit the line "However, these two sets of output do not share common start times" to add ", differing by 1 day", to remind the reader. Then, in L103, right after the line ending "sorting the runs by start month.", I would add a one or two sentence justification here for the choice of approach. E.g. "Differences in transient weather systems between the two initialisation times are not negligible, making the raw difference between GEPS6 and GEPS5 hard to interpret. Taking the difference with ERA5 effectively removes much of the variance added from the differing transients, thus allowing for a better comparison. We have documented details in the Supplementary Text." Or similar.

   Then simply remove L128. The sentence there "The usage of ERA5 to handle the 1-day difference in start dates in the formulation of (4)" does not make sense as written anyway, and it reads better to put the intuitive idea up front rather.

2. Regarding Barsugli and Battisti (1998), hereafter BB98, the Figure you showed uses the "standard parameters" for their model, which I understand to be chosen to be 'generic', or 'typical' of the midlatitudes. As Barsugli and Battisti discuss, there can be particular situations where the coupling is much stronger. In such situations, the impact of coupling versus no-coupling can become visible on shorter timescales. For example, if you double the constant determining the effect of SSTs on T2M in their model, you can verify numerically that the T2M variance is systematically lower already on daily timescales. Here is my own numerical analysis of this set-up:

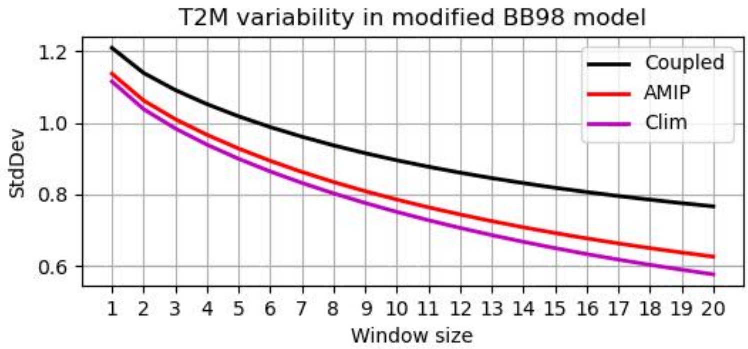

Figure: The standard deviation of T2M, smoothed using a running mean with a given window length, as a function of window length. T2M is derived from a 50,000 timestep integration of the BB98 model, where the coupling strength has been doubled. The BB98 model parameters have all been normalised by multiplying by 10^6 in these integrations. The AMIP and Clim simulations are obtained as in BB98 (prescribed SSTs and fixed climatological SSTs respectively).

Note that it's true that the same plot run using the "standard parameters" show no real difference until the window length is >20 or so, consistent with it being a low-frequency phenomenon in general. However, it seems to me that there might be particular situations where coupling is much stronger, and in these cases the effect on the temperatures, and hence fluxes, could be much larger, and in particular already visible within the 14-day timescale you are looking at.

I'm not going to make a strong claim here, because more analysis would be needed to check if this approach really makes sense. But it makes me think it is a mistake to assume that BB98 is simply not relevant for your results.

I leave it to the authors to decide what to do. My suggestion would be to at least revise to something like "Second, there is a need for more physical understanding of how two-way coupling produces better air–sea fluxes, perhaps following the spirit of Barsugli and Battisti (1998)". But I am not insisting.