This is an important manuscript that describes the changes made between UKESM1 and UKESM1.1 in order to improve the simulation of the historical surface temperature in the second half of the 20th century. A number of changes and bug fixes were made, but the key change appears to be a reduction in the magnitude of the aerosol ERF as a result of a reduction is sulphate.

The discussion focuses on one specific model but the problem of an overly cold late 20th century is present in other climate models as well. Therefore the manuscript should be relevant to a broad audience and well suited for GMD.

While I believe that the conclusions are very likely correct, thie analysis does not currently provide sufficient evidence to support them. The reported change in ERF is small difficult to attribute to aerosol only (see major comment below).

I recommend performing additional simulations (time evolving ERF calculations) to better support the conclusion that the primary reason for the improvement in surface temperature is due to a reduction in the aerosol forcing.

Major comment

The magnitude of the change in total ERF (+0.08 W/m2) appears small compared to the actual change in surface temperature (Fig 3a).

Using a simple back-of-the-envelope calculation (see Shindell 2014, doi:10.1038/nclimate2136), we can estimate the warming for a given forcing and TCR as:

dT = TCR/F2xCO2 ERF_tot

where TCR is the transient climate response, F2xCO2 the 2xCO2 forcing, ERF_tot the total anthropogenic forcing. Let's assume F2xCO2 = 3.6 W/m2 for both UKESM1 and 1.1 (based on Figure 18c) and estimate dT for both models:

dT_UKESM1 = 2.76 / 3.60 * 1.76 = 1.35 K
dt_UKESM1.1 = 2.64 / 3.60 * 1.84 = 1.35 K

Based on that simple calculation, both models would yield about the same level of warming due to a compensation between an increase in total forcing and a reduction in TCR. While that is not the case, it does make it difficult to simply conclude the all the changes arise from ERF while the TCR remains "essentially unchanged" (line 14).

Another way to look at this to calculate how large a temperature change one might expect given the change in ERF (0.08 W/m2):

dT = 2.64 / 3.60 * 0.08 = 0.06 K

This value is very small compared the actual temperature difference between the models (Fig 3a).

The most likely explanation for this discrepancy is that the difference in ERF is much larger during the period 1960-1990 than the value of 0.08 W/m2 reported for 2014.

Similarly, the forcing values presented in Table 3 are not very convincing. The total anhtropogenic forcing is indeed larger in UKESM1.1 (+1.84 W/m2) than UKESM1 (+1.76 W/m2). However, summing the components yields a smaller forcing for UKESM1.1 (+1.61 W/m2) than UKESM1 (+1.65 W/m2), and both of them are off by more than the difference between models. I don't think this data supports the conclusion that the change is aerosol forcing is key. Having comparable values for the period 1960-1990 would likely help.

I would recommend to perform additional simulations to estimate ERF for different periods more relevant to the cold bias. The best would be to follow RFMIP experiments for diagnosing time-evolving ERF. Due to the need for additional simulations, I recommend that the manuscript be returned for major revisions.

Minor comments

Lines 122-127: paragraph requires clarification. If I understand correctly, r_c was set to 10 sm-1 in GC3.1, then mistakenly to 148.9 sm-1 in UKESM1 and then to 1 sm-1 in UKESM1.1. Clarify the motivation for using 1 sm-1 instead of 10 as in GC3.1? Insufficient SO2 dry deposition?

Lines 207-211: What is the impact on net TOA radiation?

Line 271: any reason for stopping at 462 years and not the recommended 500 years for DECK piControl?

Lines 275-276: "later period". Chosen because of the smaller drift or for other reason?

Figure 3: HadCRUT5 reports SST over ice-free ocean, and surface air temperature over land and ice covered ocean. Was the same calculation done for the model output?

Line 319: 1900 → 1901 for consistency with the figures. Similar on line 324.

Lines 355-356: explain how globally averaged Nd and r_eff were calculated.

Figure 6: how different are the starting values?

Lines 365-372: are Nd anomalies really relevant if the clear-sky OSW anomalies are driving the surface temperature change?

Figure 9: explain how vertically averaged Nd was calculated.

Line 418: is the detrending actually needed? piControl looks stable in Figure 1b.

Line 425: “relatively large climate sensitivity” here, and “outside of the CMIP6 5-95% ranges” on line 552.

Table 4: use a consistent number of decimals and verify that the net adds up, or explain why.

Line 602: “thesimulated” → “the simulated”

Lines 610-611: Table 3 currently doesn’t support this assertion. See major comment above.