Review by Aaron Donohoe

This manuscript employs a single column isotropic shortwave radiation model to decompose the changes in the net surface shortwave flux in response to solar radiation management in the geomip model ensemble. The use of the single column model in conjunction with the assumption that changes in clear sky reflection and absorption are due to sulfate aerosol forcing and water vapor feedbacks respectively is very clever (especially putting these changes back into the full sky equations). However, I do question whether the cloud feedback can be isolated from the effective radiative forcing of aerosols associated with the direct and rapid response of clouds. I suggest an improved methodology below. I highly suspect that much of what the Authors interpret as a cloud feedback (i.e. associated with temperature changes) is actually the cloud changes due to the aerosol forcing itself and is better characterized as a forcing. I also question the use of the surface radiative budget as opposed to the top of atmosphere of tropopause. As such, I think the main conclusions of the manuscript are not supported and the work could be misleading for the field. I do recognize that the analysis pursued could allow the authors to determine the magnitude of forcing and feedbacks associated with each cloud, water vapor and surface albedo changes and, potentially informs which physical processes determine both the robust changes in the ensemble average and the cause of inter-model differences. There is great potential for the work to offer new insights into the response to geoengineering but, as is, the methodology is flawed and conclusions are misleading. I do not recommend publication of the manuscript in its current form; the Author's need to fundamentally modify the methodology and focus of the manuscript.

I'm not sure I understand the rationale/agree with the premise that the net shortwave flux at the surface is a useful metric for understanding inter-model differences in the response to solar radiation management (SRM). Why favor this metric over the forcing, or the net (longwave plus shortwave) radiative change either at the surface or (preferably) the tropopause? Is the an a priori physical reason to expect the correlation between net surface shortwave and temperature response? I could not find one in the manuscript. In particular, the shortwave water vapor feedback differs in both sign and magnitude when considering the surface fluxes versus the tropopause or TOA and it's hard to justify the interpretation of this feedback defined at the surface (as pursued in the current manuscript); in a warmer planet, the moister atmosphere directly absorbs more solar radiation which has a heating impact on the climate system but this reduces the downwelling shortwave flux to the surface which the Authors would interpret as a cooling feedback in the framework used within the manuscript. This feedback is found in the current manuscript to have a magnitude of order one half the net surface shortwave change and likely confuses the results and interpretation of the manuscript. I'm not sure that the correlation found between the temperature response and net shortwave flux at the surface is anything more than a statistical coincidence (given the number of independent data points available when accounting for expected correlations between ensemble members of the same model). I believe that looking at the same diagnostics (including LW changes) from the perspective of the TOA radiation alongside the surface would help to illuminate the underlying physical mechanisms responsible for the inter-model differences in the response to SRM.

Main points:

Separation of cloud feedbacks from direct aerosol forcing of clouds

Clouds respond directly to forcing agents (e.g. aerosol, carbon dioxide, etc) and to changes in surface temperature. The IPCC (and field as a whole) includes the rapid cloud response to forcing agents in the "effective" radiative forcing whereas the cloud radiative changes due to surface temperature changes are generally classified as a radiative feedback. The present manuscript associates all the cloud changes with the feedback (equation 11) and I suspect much of what is called a cloud feedback is actually inter-model differences in the effective cloud forcing. This suspicion is based on two lines of evidence:

1. The cloud radiative changes in figure 4 seem to coincide with the nearly step function changes in aerosol as opposed to the surface temperature changes. Panels E and C are the best examples. The cloud radiative changes ramp up almost immediately at 2020, before the surface temperature has decreased and return to near their unperturbed value almost immediately when the SRM stops at year 2070 even though the surface temperature takes longer to recover.
2. The published cloud feedbacks differ in sign and magnitude from those found elsewhere in the literature for the same models. More fundamentally, the Authors conclude that cloud changes damp the response to geo-engineering whereas the models included in the study have been found to have positive net cloud feedbacks in response to CO2 (see Table 1 of Andrews et al. 2012 – Forcing. Feedbacks and climate sensitivity in the CMIP5 coupled atmosphere-ocean climate models) The comparison I'm making is unfair to Authors since I am comparing net cloud radiative impacts at the TOA to the surface SW impact. However, figure 3 of the above manuscript suggests a sign difference for at least the hadGEM3-ES model. Either way, the ensemble average negative cloud feedback suggested by the Authors seems at odds with the literature, is likely confused with the effective forcing and should be further analyzed (remove forcing, look at net radiative impact, compare TOA and surface) since this result contradicts and confuses the existing literature.

A fairly straightforward solution to the above objections would be to compute the same fields outlined in equations 10-12 for each year of the simulation where the SRM is approximately constant (2025-2070 ish) and plot the radiative changes of each term versus the surface temperature change for all. As suggested by Gregory, the feedback is the slope of the linear best fit line and the effective forcing of each term is the y-intercept. This would also allow the Authors to calculate the impact of the aerosols on the shortwave absorption within the atmosphere which is alluded to in the discussion. I think this would appropriately isolate the effective forcing of clouds and the Authors might find the very interesting result that the inter-model differences in climate response to SRM is well correlated with effective forcing where the latter includes both the direct forcing of the aerosols and the rapid impact of the aerosols on the cloud radiative effect.

Use of the surface radiation budget

The surface energy budget is not closed with respect to the radiation and it is widely recognized that changes in surface radiation are balanced by turbulent energy fluxes with only small temperature adjustments. Generally, the radiative changes are viewed at a level where the system is closed with respect to radiation – either the tropopause or TOA. It is fair to challenge this paradigm and the surface radiative budget may be useful for geo-engineering but that point should be discussed and analyzed, not

taken for granted as it is in the current manuscript. In particular, one place the surface radiative changes are less than useful is the interpretation of atmospheric solar absorption on the surface energy budget. As the atmosphere warms and moistens it absorbs more shortwave radiation that would have otherwise mostly (since the majority of the Earth's surface is dark) been absorbed at the surface. As a result, less shortwave is fluxed to the surface, which would be seen as a cooling influence on the surface. Yet, in the column average, slightly more shortwave is absorbed. Since most of this additional shortwave absorption occurs in the lower troposphere, where water vapor is abundant, it is tightly coupled to the surface energy budget and will warm the surface even if the surface shortwave flux is reduced as a result. Radiative kernels estimate this feedback to result in +1.0 W m^-2 K^-1 more absorption in the atmospheric column and +0.3 W m^-2 K^-1 as measured at the TOA (Donohoe et al. 2014, Shortwavbe and longwave contributions to global warming under increasing CO2, PNAS). Therefore, the surface feedback would be deduced to be -0.7 W m^-2 K^-1 with the wrong sign and more than twice the magnitude of the changes at the TOA. In the very least, the manuscript should include similar diagnostics at the TOA to resolve this sign paradox and a discussion of these points to support the assertion that surface shortwave changes are a useful metric.

To play devil's advocate, it seems like most of correlation between the temperature response and net surface shortwave comes from the forcing. Is the use of net shortwave at the surface a better predictor of the temperature (statistically distinguishable) from that of forcing alone (surface or TOA)? The latter certainly would result in a stronger regression – and one more consistent with climate sensitivity—than using surface shortwave even if the correlation is slightly worse. More generally, what would the correlation be if one used forcing alongside published estimates of the model's climate sensitivity in response to CO2? It looks like the outlier from the strong relationship between forcing and response is the MIROC-CHEM-AMP which has a pronounced cloud feedback. Is that cloud feedback part of the climate sensitivity of the model or (I would guess) the direct cloud response to aerosols? The cloud changes seem to ramp up alongside the forcing in figure 4f. I elaborate on the interpretation of the cloud feedback term below.