General comments:

This manuscript is a significant contribution to the field of radiative feedback and radiative adjustments studies where radiative kernels are frequently used but, as the authors note, not always readily accessible. This has the potential to advance modelling science of particularly radiative feedback studies, but also studies of radiative adjustments. The tool itself should be very useful for scientists in the field, and while the analysis presented comparing different kernels is not extensive, it is sufficient to illustrate some interesting novel results and could easily lead to further investigation. The manuscript is well written, with the methods used mostly very clear and the results clearly support the interpretations and conclusions made. A few points would however, benefit from clarification with additional information or rewording. The structure of the manuscript is clear and concise, the manuscript reads well, and the abstract and title describe the contents of the manuscript well. However, in a number of places the language describing results is vague and would benefit from the use of actual quantified values (most of which could be derived from the figures). Including values in the conclusion and abstract would really help to emphasise the key results and highlight the interesting results the authors have found to readers. Other papers are properly referenced and only one or two points are missing a reference. The supplementary material documenting the code and datasets is extensive and should be very useful to potential users. Overall the manuscript is a great contribution to the literature and I suggest only minor amendments - some to clarify the method and results, and others to maximise its usefulness to potential users - and can only wish such a tool were available when I first became interested in radiative kernels!

Specific comments:

- There are a number of instances in the paper, particularly the results, where results are described only qualitatively. Generally it would be better to give quantified values in text, which also describe the sign of the change, rather than words like 'highly' or 'considerably' which are subjective and sign agnostic. While readers can see results in the figures, using actual values in the text would help make the key results and their significance clearer to the reader. I've noted more examples in the line by line comments below, but for example in line 324, you could say that 'the clear sky kernel is up to 1 W m^-2 more negative than the all-sky kernel', rather than just 'the surface temperature kernel is highly sensitive to clouds'
- While table 1 offers some useful comparison of the differences between the kernels, there are a number of factors not included, some of which may be of greater importance. Firstly, the number of years averaged over to generate the kernel would seem as important as number of levels and resolution. Likewise, whether the kernel was computed with aerosol included or not (for example, the HadGEM3 kernel was computed without aerosol and so effectively represents a 'clean-sky' kernel; whereas the CAM5 and ERA5 kernels were computed with aerosol represented in the radiation code) could have a big effect and is worth stating. Other differences may be less important, but perhaps could still be included as columns, such as the reference climate state used (e.g. pre-industrial or present-day), and the data source used (e.g. model, reanalysis, satellite). While readers could find these details from reading the referenced papers, it would seem very useful to readers and users of the code to include a few more of these differences here.
- On lines 172-173 and 239-240, if I understand correctly, the functions calculate a monthly mean climatology from the control simulation input and subtract from the un-averaged perturbed simulation input? Is it expected that the input from the perturbed simulation will already be given as monthly means too? If not, then what is the reason for taking monthly climatology of the control experiment input, but differencing to non-monthly meaned perturbed experiment input? It may be the sentences just need re-wording.

- On line 208, what does 'masking below the surface' mean? Presumably this is related to how surface/orography is defined differently among kernels and models, but it would be good to explain this, and perhaps explain somewhere how these differences are dealt with. And state if this is also applied to the other kernels not just the water vapour kernel.
- It would be great to evidence and emphasise the second conclusion point by stating what is the largest difference between two kernels in terms of the total climate feedback, or simply the stdev of the sum of feedbacks across the kernels. This would also be a great 'headline' result to show readers in the abstract and help quantify the text in lines 13-14. (Perhaps going further, you could even express this in terms of the uncertainty it would add to the ECS calculated from the sample input, but that might add too many factors, so is just an idea).
- On lines 414-415, while a discussion around the validity of multi-model-mean approaches is much bigger than needs addressing here, there are issues with encouraging users to use the mean of all the kernels as a simple way to eliminate biases. For example, clearly some kernels are not independent, such as CAM3 and CAM5 or HadGEM2 and HadGEM3 kernels and could share similar biases. Perhaps you could re-word to suggest that using the mean of several kernels that the user selects (in addition to using multiple kernels for sensitivity analysis as suggested in your previous sentence) might be better than a single model, but the user should decide which ones to include.
- Not essential but more of a question and observation: lines 380-381, since the spread in the SW cloud feedback is influenced by the spread in the surface albedo feedback, would you expect the relative bias of each kernel to the mean to be somewhat inverted between the two feedbacks? It looks like one does see that to a degree: CAM3 and GFDL are the highest over the Arctic for SW cloud but lowest for albedo, whereas the reverse is true for BMRC and CAM5 (excluding HadGEM2, which perhaps has some other differences with the SW cloud kernel). It would be nice to mention this if you agree. And also perhaps suggest an explanation why the HadGEM2 kernel is such an outlier for SW cloud feedback
- I noticed that the ECHAM5 kernel is included in the kernel data repository alongside the 11 mentioned in the paper - is there a reason this was included in the repository data but not in the analysis in the paper?

Technical corrections

- 11 - would be great to specify here the sample climate model output is representing a 2x CO_2 perturbation
- 13-14 - would be nice to include some quantification of the spread - would highlight the significance of the results to readers
- 65 - quantify how much the surface albedo kernel varied by
- 66 - again ideally quantify
- 77-79 - need reference(s) for these examples
- 146 - Smith et al 2018 may not be the first ref for this - other studies used kernels for adjustments earlier (e.g. Vial et al 2013, Block and Mauritsen 2013)
- 160 - define 'a control and a perturbed simulation' - it might not be immediately obvious to all readers
- 188 - 'holding the control and perturbed simulations' - should this be 'holding the input data of the control and perturbed simulations'?
- 241-242 - is the option to calculate all-sky or clear-sky feedbacks also applicable to other functions? If so, it would be worth mentioning it earlier in the first function which it is true for, or in the intro to section 3

- 250-252 - while its great the kernels can be used for radiative adjustments too, I am not sure the cloud kernel will work for radiative adjustments with either the adjustment or residual method equations given in section 3.4 - this may be worth noting
- 306 - strictly speaking, these are minima for the means, being negative not positive, but perhaps could be described as 'maxima in magnitude'
- 309-310 - would be good to suggest an explanation of how the inclusion of clouds or not might cause such a change in the spatial pattern of the mean temperature kernel
- Figure 1 and figure 2 both skip the subfigure label 'g' and go straight to 'h'
- 320 - avoid use of 'significantly' where not referring to statistical significance; 'larger' would suffice, or better yet quantify it
- 329 - The numbering of the section would make more sense as 4.2 here rather than 4.1.1 (as there is presently no section 4.2, nor a 4.1.2 either)
- 333-334 - explain whether using the last 30 years of a 150 year simulation means that the surface T response has equilibrated sufficiently, or use a reference to state this is a sufficiently long enough length of time to use
- 363-364 - The description of figure 2j is inaccurate: the maxima over NE Pacific and the southern Atlantic and Indian Oceans are similarly large to the two mentioned in the text, and the contrast between land and sea is not as clear as between different land areas or different sea areas. I would suggest a different description.
- Figure 2 - it would be helpful to separate the cloud feedback into LW and SW components, to match the treatment as in figure 3. This should help with the description of the spatial pattern in the text too. Perhaps this could be achieved without reducing the subfigure size (and still fitting on one page) by making it 4 columns instead of 2 columns of subfigures.
- 372 - 'very large' - please quantify (especially because it is an interesting result!)
- 404-405 - this is inconsistent with line 328, where you noted that the surface albedo kernel spread <u>is</u> indeed different between the all and clear sky in the tropics; I'm also not sure that this would prove the point that clouds are not the dominant cause of spread across kernels. I do agree however, in figure 1, that only the T kernel spread seems to differ a lot between all and clear sky - so maybe re-word this sentence if you want to make this point, or remove it.
- 408-410 - I am not clear how the global mean feedbacks in table 2 support the interpretation that the lapse rate or surface albedo feedbacks are most important for Arctic amplification?