West et al.: Using Arctic ice mass balance buoys for evaluation of modelled ice energy fluxes

The authors use ice mass balance buoys (IMB) to estimate fluxes through the top and bottom of sea ice. The authors present this new method and then compare the observed fluxes in the North Pole and Beaufort Sea regions. The authors then compare the observed fluxes with modeled fluxes from the HadGEM2-ES climate model. The main findings are that there are biases in these fluxes in the model, which are likely due to the biased mean state of the model. Additionally, there are differences in the fluxes in the Beaufort Sea and North Pole regions. I have a few major and moderate concerns about the way the model and observations are compared and these need to be addressed before I can recommend publication.

Major concerns:

1) Internal Climate Variability

   The elephant in the room for a comparison between a climate model and observational data is the issue of internal climate variability (see references below), which you never mention, and leads me to have major concerns with your method. This is also relevant for when the Arctic will become ice free, which you mention in the introduction. Since HadGEM2-ES is a fully-coupled, freely-evolving climate model a single model experiment should not be expected to match the observed sea ice conditions. You do not mention using ensembles and where/how the observations fit in an ensemble spread.

   Indeed, it appears you are comparing the mean climate model state with the mean observations (Fig. 8). This is not a particularly useful comparison – we know the model is biased from your previous work and therefore we expect to see biases in these mean fluxes as a result! Instead, a more useful analysis would be to evaluate situations when the model does have similar thicknesses to those observed, do the fluxes match the observations? That would tell us more about the processes going on in the model and how well they compare to those observed. You could do this by plotting the distribution of the conductive fluxes by thickness for the model and observations for a particular month or throughout the year.

   Two additional comments on the model: a) It would be useful to quantify other relevant model biases to the sea ice mass budget like SST or ocean heat transport; b) You compare different years from the model (1980-1999) and observations (1997-2016). I know you did analysis about how the periods are different (Pg.12, lines 13-24) but why not just use the same years that presumably have comparable radiative forcing?

   Kay et al. 2011 (doi: 10.1029/2011GL048008)

   Kay et al. 2015 (doi: 10.1175/BAMS-D-13-00255.1)

   Jahn et al. 2016 (doi: 10.1002/2016GL070067)

   England et al. 2019 (10.1175/JCLI-D-18-0864.1)

2) Additional sources of uncertainty

You mention uncertainty in salinity as one of the big uncertainties in the IMB flux calculations. I think you need to mention that there are also large ranges in the observed snow and ice densities (see refs below) that could cause uncertainty in the retrievals. The values you use are reasonable, but you need to at least acknowledge this and do some basic calculations about how big a difference these values make.

Franz et al. 2019 (doi: 10.5194/tc-13-775-2019)

Webster et al. 2018 (doi: 10.1038/s41558-018-0286-7)

3) Significance

You spend much of section 3 describing differences in observations in two regions and sections 4.1/4.2 describing differences in the mean state of the model and observations. No significance tests were discussed to indicate whether the means are really significantly different in Figs. 7/8. A simple t-test should suffice, but without this I have a hard time believing some of the conclusions (e.g. that September fluxes differ in the IMB between regions).

Moderate concerns:

1) Model thermodynamics

I am surprised that HadGEM2-ES, a CMIP class climate model, uses the very simple zero-layer thermodynamics and I think that this should be addressed. The Bitz and Lipscomb thermodynamics is more realistic than zero-layer, and even this has been superseded by the mushy layer thermodynamics of Turner and Hunke (see below). I realize you can't change the model at this point and this shouldn't prevent publication, but your own results show that the assumptions of the zero-layer scheme for conductive fluxes are bad (Fig. 7 and Table 1). Maybe one of the conclusions should be that the zero-layer cannot represent the observed processes so HadGEM2-ES might want to stop using it?

Bitz and Lipscomb 1999 (doi: 10.1029/1999JC900100)

Turner and Hunke 2015 (doi: 10.1002/2014JC010358)

2) Figures

Individually these comments are fairly minor, but the sum of them is moderate.

- Table 1 – Please add the # values per month and per flux to this table rather than just listing them in the text. Also, adding the units below the flux names (not just in the table description) would be helpful. I expect top melt to usually be in cm/day (or kg m-2 s-1), not W/m2.
- Fig. 1 – It would be helpful to add the sign convention for fluxes here at the interfaces (show the flux direction for positive!). Either label or remove the red/yellow arrows.
- Fig.3 and Fig.4 – you don't have units on the y-axis or in the labels.
- Fig. 4 – Please define surface_r and interface_r. Is snow depth the difference between these two lines? Please clarify on the figure and in text.
- Fig. 5 – I think a diagram of an IMB would be helpful for modelers, which I think this one is, but it's poorly labeled. What do the dots represent (thermistors? Is this why they go above the snow interface)? What does the L/R position of the dots mean (temperature?)? Why is there a green line over part of the dots?

- Fig. 6 – the circle colors on this figure are very hard to make distinguish. Perhaps different shaped symbols in black would be better? Again, define surface_r, interface_r, and bottom_r. Are points where there is blue (aka surface_r) mean that the blue and green circles are overlapping? Adding arrows to indicate the transitions for the false bottom would also be helpful? This figure also makes me question why you don't linearly interpolate between the "correct" depth – it looks possible so why lose that data?
- Fig. 7 – It might be clearer to show the total spread in values with shading and then just a central dot for the mean since the individual points with their spread get hard to distinguish. Also, in text you list the means but they need to be shown to be significantly different.
- Fig. 8 – What purpose does this figure provide other than the model bias, which we already expected from your previous work. Again, if you do show it, significance tests are important. I think a PDF of the flux by thickness would be more helpful to supplement this figure.

3) Code availability
   The effort put in here by the authors to make the data available is lackluster. I understand the model code itself may not be available. However, the authors should make more effort to list how to get the buoy data (raw and processed) as well as the model data (if the code isn't available) since these should be public if it's part of the CMIP archive. See the guidance on the website: https://www.geoscientific-model-development.net/for_authors/code_and_data_policy.html

Minor concerns:

- Shu et al. 2015 isn't in your references.
- Pg.4, line 6-8 – do you mean the ice temperature or air temperature? What's going on to cause the non-physical temperatures?
- Pg.4 line 14 - How long do buoys last? Give a range here.
- On page 7 you state the convention of positive fluxes indicate downwards. It would be very helpful if you mention that much earlier on Page 4 line 21. And put this on a diagram (Fig.1 and/or Fig.5) too.
- Pg.7 line 20 – how thin is "quite thin"? Be specific!
- Pg.7 line 26-28 – These means are over all available buoys and all years, right? It would be good to be explicit.
- Pg.9 line 15 – change "7 to 143" W/m2 because negative fluxes are possible and the current wording is unclear.
- Pg.8 line 25 AND Pg.10 line 1, mention that those regions are defined in Fig.2.
- Pg.12, line 5 – what is the model grid (it hasn't been mentioned)? Why the huge range in grid cell size?