Response to Review of Watson-Parris et al (2024)

Firstly, we would like to thank both reviewers for their thorough and considered feedback on our paper which has helped to improve and clarify some key aspects. For context we include the full reviews below and reply to each comment in turn, making comments in blue and highlighting particular changes to the revised manuscript in orange.

Reviewer #1

The years 2023 and 2024 have seen exceptional levels of global warming, and climate scientists are still trying to attribute this warming to different causes. Watson-Parris et al. add to the growing body of literature assessing the impact of IMO shipping regulations on near-term warming.

They not only present a new set of simulations which is broadly in agreement with previous work on this topic, but also try to reconcile different climate model results which seemingly came to different conclusions. Their discussion on methodology and framing is very valuable. I think this is an excellent piece of work, and would like to see it published after the authors consider the minor comments below.

We thank the reviewer for the positive and constructive feedback on our manuscript.

Comments

Title **:** Avoid having a subjective assessment of the strength of the IMO effect in the title ("Weak"). Maybe saying "within internal variability" is enough and more objective. Assuming the radiative forcing is 0.1W/m2 or more, that corresponds to ~1/10th of the aerosol ERF which is not insignificant. How that translates to surface temperature change may be considered a different question. There could be competing effects: it takes time for the surface to warm and hence for the signal to emerge, and the background scenario also has declining shipping emissions so the difference between control and perturbation scenarios gets smaller after 2030. This probably explains why the temperature change maximises around 2030 in various models as discussed around L.288.

Thanks for the suggestion, I like this rephrasing which also avoids the slightly awkward qualification. We've changed the title accordingly:

> Surface temperature effects of recent reductions in shipping SO_2 emissions are within internal variability

Have you compared your TOA SW and LW radiation to observations as in Quaglia & Visioni? Their figure 1 shows a very large increase in ASR at the TOA in 2023 which is not captured by the models (with and without shipping regulations, though they do help). Of course this increase can come from other aerosol changes or natural variability. Maybe the observed

changes are within the variability of your larger ensemble? If it isn't, does that say something about the ability of CESM2 to capture observed changes (especially given large uncertainties around aerosol ERF)? **I recognise this may be a lot of extra work and would be happy to see this work published without it.**
We thank the reviewer for this suggestion, which we agree would be an interesting investigation. As the reviewer points out though, this would be a lot of extra work that we feel would be better suited to a separate paper. Indeed, many of the authors on this paper are involved in a more targeted follow-up paper which utilizes additional simulations from RAMIP.

L.64-5 What's the reference for the 0.5W/m2 value? Diamond (2023) gives a value of order 1 W/m2.
This value is the actual value from Diamond (2023), rather than the approximation given in their abstract. From their paper: "The IMO 2020 regulations led to a ∼ 2 W m−2 IRFACI within the shipping corridor during austral spring and a ∼ 0.5 W m−2 IRFACI in the annual mean." We have added the citation to the text:

> ...estimated to be 0.5 W m^{-2} in the annual mean within shipping corridors (Diamond, 2023).

L.92 What Quaglia & Visioni call a radiative forcing is in fact a change in TOA radiation between 2 sets of coupled simulations, rather than being a ERF calculation which assumes fixed SSTs to remove the effect of feedbacks. So change "ERF" in L. 92.
Good point, thanks for catching this. We've updated L92 to:

> ...approximately 0.2 W m^{-2} radiative perturbation in CESM2 (for a 90% reduction in shipping emissions).

L.283 the p=0.18 is for 2020-40 I imagine?
Yes, this has now been clarified in the text:

> However, our 18-member ensemble shows that the global ensemble mean warming over 2020-2040 is not statistically significantly different from zero (p=0.18), even in 2030 (p=0.054)

Make the figures higher resolution (plt.savefig('filename.png', dpi=300) for ex in matplotlib).
The figures will be updated with higher DPI in the final version, thanks for spotting this.

Reviewer #2:

In this paper, the authors use a 18-members ensemble of CESM simulations to investigate the potential impacts of the changes in shipping emissions in 2020. The paper has a rather comprehensive review of recent papers on the same subject, and is definitely well written. I agree with the first reviewer about the value of their discussion and the importance of this study.
We thank the reviewer for this positive and constructive review.

I also agree with the first reviewer on the framing problem: "weak" is a personal characterization of the results, and one that I'm not sure is very informative here. It is clear, as the authors also note, that this signal is not so easy to detect, and different investigation methods yielding different results speaks volumes to that. However, the authors also say that "IMO regulations may contribute up to at 0.16 ℃" (note the at typo in the abstract) for individual years, and I think it's hard to reconcile this presumpt "weakness" with the relevance that that might have to, for instance, intensifying heatwaves or fire weather over specific regions.

Yes, we acknowledge that 'weakness' is somewhat subjective and perhaps unhelpful. We have now rephrased the title based on Reviewer 1's suggestion to:

> Surface temperature effects of recent reductions in shipping SO_2 emissions are within internal variability

Differences in conclusions are particularly interesting when considering that, using a very similar set-up (with the only difference being the magnitude of the SO_2 reduction, 80% and 90%, and the size of the ensemble, 18 and 10), their results are strikingly different from those in Quaglia and Visioni (2024) (Q&V), in review for ESD. In these results, I wasn't able to find any mention of what size of ensemble the authors here use for the baseline simulations: do you use the whole 50 (considering only the smbb runs) or 100 CESM-LE, or only the same ensemble members you spun-up from?

Yes, this is an important value to quote. We compare against the 18 unperturbed simulations, which should be closest in state to the perturbations and have added a clarification in the methodology:

> All comparisons and differences are calculated with respect to the 18 corresponding unperturbed ensemble members.

I think this could be useful to understand why (acknowledging they might be using slightly different methodologies) they find no significance in Fig. 3 while Q&V find statistically significant differences both in TOA fluxes and in detrended monthly temperatures, as well as the maps in their Fig. 3. Could it be that a 10% further reduction contributes to the significant/not-significant threshold? Maybe the two teams can find a way to compare their results, either here or in the future.

We agree this is an important distinction, but as we discuss in the paragraph starting L285, we believe this is more a question of framing than methodology. We would be happy to share our data with Quaglia and Visioni if they are interested in performing a more detailed comparison of our results. However, we would prefer to keep a broad discussion of the literature in this manuscript, rather than focus on the comparison with one previous study, as the importance of framing for the conclusions drawn from different analyses is an important point to make.

One possible curiosity would be to check if all the simulations were run on the same cluster as the original CESM-LE runs: in the Acknowledgments here, it looks like these

simulations might have been ran on another one? If there is no bit-by-bit reproducibility in the branch, this might influence such a delicate assessment.
While 8 of the simulations were simulated on the same cluster (and we believe the same cluster as for Q&V), another 10 were simulated elsewhere. We believe that a comparison between 18x 20 year fully coupled simulations should not be affected by machine differences, especially as we consider time means, rather than transient evolutions.

I also think the maps here in Fig. 3 pick two specific cases in which the results w*ill* look less significant by design: for the 2020-2025 period, by including a at least 12 months period in which the "termination shock" (to quote from the Yuan et al. (2024) paper) hasn't manifested yet, while for the 2020-2040 period, by including years in which the scenarios somewhat reconcile in terms of shipping emissions.
We feel these are the appropriate time periods to answer the questions at hand: what was the contribution of these emissions changes to the record temperatures in 2023; and what is the climatic impact on global temperatures? We agree that we could find significant regions / time periods but the very act of having to refine our search in such a way demonstrates the regionality of the impacts.

I want to stress that I'm not trying to imply that one study is wrong and the other right: I think that the fundamental differences in answer come from differences in methodology and perspectives: this study takes a more "climatological" approach and find that, especially when considering the underlying global warming trend, the signal is hard to detect over decadal timescales, whereas Q&V take a more "detection and attribution" approach and ask if, for a specific year, close enough to the termination effect, it is possible to make a probabilistic determination of how even a "minor' effect might have contributed to pushing even further a year that is now almost universally recognized as highly anomalous. Indeed, I think in the abstract itself, as I pointed above, the authors acknowledge that they can't exclude that IMO changes have produced a significant change in the specific year under question. So I strongly suggest the reviewers reconsider some of their language in light of this.
We appreciate your frank remarks. We agree, the differences are mostly around a question of framing, and you are right that we take a more climatological framing as opposed to the D/A framing in Q&V. We tried to highlight this in the discussion. We try to be as objective as possible in the discussion, particularly in the comparisons with Q&V which we believe are fair and balanced. As suggested, we have updated the title to make the headline assessment less subjective.

Minor comments:
Sometimes the authors use ºC, sometimes they use K, even in the same phrase (see line 77). I suggest to reconcile that.
Yes, thank you. We've switched to consistently use ºC.

Note the "at" added to line 33 of the Abstract.

Fixed.

Fig. 4: are these results implying that during a positive ENSO phase, in 2023-2034 one would have expected a cooling contribution from the change? I think this is also rather different from the Q&V analyses, and very counter-intuitive...
This is not what is implied by Figure 4. This Figure merely highlights that the response to the IMO regulations is not influenced by the ENSO state in the year in which the IMO regulations came into effect: there are no significant differences between three lines. This is a distinct analysis from that performed in Q&V who rather showed the change in the Nino3.4 index under the emissions perturbation, whereas in Fig 4 we show the change in global mean temperature for simulations delineated by their (unperturbed) phase in 2020.

Fig. 5: There seems to be a very weird ensemble member that starts at 1.5 and cools down to 0.9 by 2025. That sounds highly anomalous.
We were also surprised by the high inter-annual variability observed in LE, which motivated us including this figure in the paper. We have checked our analysis code, which simply takes the global mean of each ensemble member and plots the time-series and can find no errors. We believe this speaks to the large internal variability of the system and the challenge of attributing a single hot year as a forced response.