

Responses to reviewer comments on:

<https://doi.org/10.5194/egusphere-2026-2654>

We thank the reviewers for their time spent on the paper and their detailed comments and suggestions to improve the manuscript.

In response to reviewers' comments, we have made some significant changes to the manuscript. Of particular relevance is an analysis of aerosol size distributions derived from the ACRUISE observational campaign (Yu et al., 2020). This clearly shows that the aerosol size distribution on a regional basis is dominated by the Aitken mode, consistent with earlier observations (Hobbs et al., 2000; Petzold et al., 2008; Jonsson et al., 2011). This analysis also shows that the modelled size distributions in our updated experiments with reduced primary sulphate size and increased mass fraction are closer to the observed aerosol size distribution than the default model parameter experiment. More in depth analysis comparing our results to satellite-derived constraints on the cloud microphysical response and resulting shortwave forcing have also been added to the manuscript. Comparison with the cloud microphysical and radiative response to IMO2020 reported by Diamond et al. (2023) shows that our highest-ERF experiment, Ait_10, overestimates the observed response, whereas Ait_High_4.5 (ERF 0.31 W m^{-2}) does not, and agrees more closely with the satellite-derived estimates than Base. We therefore now favour 0.31 W m^{-2} over the 0.41 W m^{-2} upper bound which was more central in the original manuscripts discussion. Taken together, these additions provide substantial new observational evidence that the aerosol microphysical assumptions currently adopted in global climate models are inconsistent with observations.

A full response to reviewers' comments is given below. All references to section, figure and line numbers refer to the revised manuscript unless explicitly stated otherwise.

Black = reviewer comment

Blue = responses and updated text

Green = original text in the submitted manuscript

RC1

The authors present a detailed study using the UKESM1.1 climate model to evaluate the sensitivity of the effective radiative forcing (ERF) from the IMO2020 shipping regulations to near-source emission plume parameterizations. They show that shifting primary sulphate mass from the default accumulation and coarse modes to the smaller Aitken mode significantly increases the modelled ERF from 0.10 W m^{-2} up to 0.31 W m^{-2} , and that increasing the mass fraction further amplifies this effect.

They conclude that recent global climate model studies converge on a tighter, lower range because they share structural biases in sub-grid plume configurations. Overall, the paper addresses an important and debated topic regarding the radiative "unmasking" effect of marine fuel regulations. It provides a good discussion around how this fits with the related literature and I enjoyed reading it. However, several major conclusions are currently overstated, and critical observational and physical constraints should be addressed before the manuscript is suitable for publication in ACP.

Responses to "Major Comments":

Comment 1:

"Single-Model Limitations and Overstated Scope: The manuscript frames its findings as a definitive exposure of a systematic, structural underestimate across all recent global climate modeling literature. However, these parameter sensitivity experiments themselves are conducted entirely within a single model architecture (UKESM1.1) running in an atmosphere-only configuration. Because individual climate models feature highly distinct background aerosol fields, boundary layer representations, and vertical velocity distributions, the claim that the entire field underestimates forcing by a factor of 2 to 4 is an overinterpretation. This framing needs to be toned down to reflect a specific model sensitivity or a structural risk within shared parameter lines, rather than an absolute systemic truth."

Response

We agree the framing needed tightening, and have revised the title, short summary, abstract, and conclusions accordingly. Our intended claim is less generic than the manuscript conveyed, and we have made several revisions to make this explicit. In the revised manuscript, the direction of the underestimate in radiative forcing is supported by several independent lines of evidence, while the magnitude is model specific but supported by prior work in independent models, and available published observational constraints.

On the claim that the consensus ERF of $\sim 0.1 \text{ W m}^{-2}$ is an underestimate, our reasoning is as follows.

(1) Our default configuration produces an ERF of 0.10 W m^{-2} , matching the multi-model mean of the estimates in the post-2020 literature, i.e. UKESM1.1 is representative rather than an outlier.

(2) The revised parameter choices are now constrained bottom-up against observed size distributions from observational studies that specifically targeted emissions from shipping in a new results Section 4.1, added in response to Comment 2.

(3) We demonstrate that multiple independent models in previous studies found a comparable increase in the shipping aerosol forcing under similar or identical primary sulphate changes in section 4.3.

(4) As shown in Section 4.4, the models used to compute the full IMO2020 ERF share the same or very similar sub-grid emission assumptions, so the mechanism we identify applies to them also, not only to UKESM1.1.

(5) The default model configuration significantly underestimates the Diamond (2023) and Zhang et al. (2025) satellite-derived constraints. This is now shown more clearly by further analysis added in response to Comments 3 and 4 (updated Figure 6, and Figure A1)).

On the magnitude, we agree with the reviewer, the factor of 2 to 4 is the sensitivity range of our UKESM1.1 experiments, dependent on many model-specific features. The submitted conclusions stated this explicitly (L333-335 of the originally submitted manuscript), but the title and short summary did not reflect it, and we have revised them as detailed below:-

Changes:

- Title changed from: "Recent Modelling Studies Systematically Underestimate the Warming from IMO2020 Shipping Regulations" to "Evidence That Recent Modelling Studies Underestimate the Warming from IMO2020 Shipping Regulations".
- The short summary has been re-written so that it no longer includes the 'factor of 2 to 4'.
- The abstract no longer includes the 'factor of 2 to 4'
- The fact that our aerosol number size distribution results, which we have added in response to comment two, are also likely model dependent has been explicitly raised in the text added in the new section 4.1.
- The conclusions have been substantially reworded to be clearer on the distinction between the claim that the forcing has been underestimated and the specific magnitude. The 'factor of 2 to 4' framing has been removed.

In addition to the above changes, further revisions made in response to comment 2 (below) also help to balance the manuscript and support the conclusions.

The sentence on L333-335 of the conclusions in the submitted manuscript:

"However the magnitude of the sensitivities identified here will likely vary between models due to differences in updraft velocities (Virtanen et al., 2025), and the community would benefit from further investigation with other CMIP6 and later generation climate models."

This statement has had more detail added and been moved to the results and discussion section 4.2, L 302-304:

“The exact magnitude of the sensitivities identified here will likely vary between models due to model-specific processes, for example differences in updraft velocities (Virtanen et al., 2025), background aerosol, sulphate lifetime (Ahsan et al., 2023) and cloud adjustment strength (Rasch et al., 2024).”

And is restated in the conclusions (sect. 4, L460-461):

“However, the exact magnitude of the ERF increases identified here is likely to vary between models (Sect. 4.1).”

The need for future work to better constrain the IMO2020 ERF is also now highlighted more prominently at the end of the conclusions (L473-476):

“For these reasons, future work targeted at better constraining the IMO2020 ERF is critical. This includes work on bridging scales to provide better integrated representations of emission plume and cloud processes in GCMs, comparable sensitivity analysis to those conducted here in other climate models and further work constraining model responses with observational constraints on IMO2020.”

Comment 2:

“Missing Bottom-Up Observational Validation: The authors rely heavily on aircraft campaign observations from ACRUISE to physically motivate their parameter adjustments (such as the shift to the Aitken mode and higher primary sulphate fractions). However, a direct bottom-up comparison with this data is completely missing. Showing that tweaking parameters improves a model's spatial correlation with top-down CERES satellite anomalies is not proof that the underlying plume microphysics are handled correctly. The paper needs to include a quantitative evaluation showing that the updated configurations (e.g., Ait or Ait_10) actually reproduce realistic size distributions and near-source properties (and ideally cloud droplet responses) when verified directly against the in-situ aircraft measurements.”

We agree this was an important limitation of the submitted manuscript. To address it, we have added a new subsection (Section 4.1) directly comparing modelled aerosol number size distributions against the ACRUISE in-situ measurements, supported by two additional experiments (Table S1) and Figure S2. The comparison shows that the Aitken-mode configurations reproduce the observed plume size distributions substantially better than the default settings. Specifically the Ait_10 experiment best reproduces the observed remote background aerosol distribution from Yu et al. (2020). As noted in our response to Comment 1, we state explicitly in this section (L239-242) that the comparison implicitly includes model-dependent representations

of plume processing. We also acknowledge there (L242-244) that the improved agreement with the observed size distributions does not rule out other contributing factors to the default model's bias, which would not necessarily imply a stronger IMO2020 forcing, and that it may not apply to other models which will have different background size distributions. The end of Section 2, the conclusions (Sect. 4.5) and abstract have also been updated to integrate this new results section.

All content of **Section 4.1** is new.

On cloud droplet responses: we have added further support of our hypothesis of underestimation of the climatic impacts of the IMO shipping regulations. Diamond (2023) (hereafter D23) reports the IMO2020 change in cloud droplet effective radius within the Southeast Atlantic shipping corridor from a counterfactual kriging analysis of MODIS retrievals. This constrains the cloud microphysical response directly, not just the top-of-atmosphere radiation, and effective radius change is a direct proxy for the N_d change. We compare their results against the effective radius response (Δr_e) in our experiments in Appendix A Figure A1, discussed in Section 4.5.1 (L391-399). We also now present the shortwave Twomey effect forcing constraint from Diamond (2023) more prominently in Figure 6 and Figure A1 (see response to comment 4). This comparison demonstrates that the model default experiment (Base) significantly underestimates the cloud droplet effective radius and shortwave twomey effect forcing response to IMO2020 compared to D23, with no overlap between the 90% confidence intervals. Several of our updated experiments, including Ait_High_4.5 with an ERF of 0.31 W m^{-2} , bring the model closer to D23's estimates without overestimation of either the annual or austral spring (SON) mean values (Figure A1).

A larger scale validation of the N_d response is not currently feasible. UKESM1.1 has a known positive N_d bias in the main stratocumulus regions (Ghosh et al., 2025, their Fig. 3), unrelated to our perturbations, so comparing background N_d against observations would conflate this bias with the perturbation response. This is why we compare changes rather than absolute values. Isolating the IMO2020 N_d signal from cloud variability at large scale is also not achievable with the current length of the post-2020 satellite record, which is why we rely on Diamond's corridor-scale constraint, designed to establish causality, as an appropriate comparison. We now highlight this in the conclusions (L468-471) as an important area for future work:

“For these reasons, future work targeted at better constraining the IMO2020 ERF is critical. This includes work on bridging scales to provide better integrated representations of emission plume and cloud processes in GCMs, comparable sensitivity analysis in other climate models and further work constraining model responses with observational constraints on IMO2020.”

The abstract and conclusions have also been updated to integrate the additional results from section 4.1 and section 4.5.

Comment 3:

“Confounding SST Feedbacks: The experiments use fixed 2005–2014 sea surface temperatures (SSTs) and sea ice configurations to isolate ERF. While this is standard practice for diagnosing uncoupled radiative forcing, it leaves the study structurally blind to alternative physical explanations. Compelling recent work has demonstrated that a substantial portion of the post-2020 global temperature surge in general and shallow cloud anomalies in the North Pacific in particular may be a response to changing SST patterns and coupled ocean-atmosphere dynamics (Zelinka et al. 2026; Tam et al. 2026; Ceppi et al. 2026). By clamping the SSTs, the model interprets these discrepancies entirely through aerosol sensitivity to match top-of-atmosphere fluxes. This is by construction, and makes sense in order to focus on the aerosol forcing, but the authors should explicitly discuss the other contribution to the observed CERES flux changes that this study explicitly doesn’t address and argue why they assume it to be small.”

We agree that this is an important point and should be highlighted in the paper, we have updated the following section to explicitly raise this limitation in our methodology and highlight the importance of coupled feedbacks which has been demonstrated in recent work:

Previous version (lines 309 and 310, section 4.4 in the pre-print):

“This means that under default emission parameters, IMO2020 cannot explain the observed regional CRE anomaly over the SEA, but with updated primary SO₄ in UKESM1.1 a clear spatial fingerprint emerges.”

New version (section 4.5.3, L421):

“This means that in Ait_10 there is a clearer spatial fingerprint of the IMO2020 forcing in the CERES record over the SEA, than under default emission parameters. Variations in SST and associated coupled cloud feedbacks are also likely to have contributed to the observed changes in CRE (Gettelman et al., 2024; Ceppi et al., 2026; Zelinka et al., 2026). These will not have been captured in our simulations that relied on fixed SST experiments to isolate the aerosol forcing. However, the spatial coherence between the observed and UKESM1.1 CRE anomalies, post-2020, indicates a substantial role for the aerosol is plausible.”

We have slightly reworded the short paragraph starting on line 311 in the pre-print (immediately following the section above) to make it clearer that our results do not

support an IMO2020 ERF of 0.5 W m^{-2} by adding ‘help reconcile’ to replace ‘reconcile’ as our results do not eliminate this discrepancy (L428-429).

We also made a small edit to the following sentence to make it clearer that Gettelman et al. (2024) also used the spatial correlations between fixed SST experiments compared to CERES CRE anomalies. We have added: ‘(from fixed SST CESM2 experiments)’ (L416).

Note, we do not claim at any point that our results demonstrate that contributions to recent CRE anomalies other than the IMO2020 aerosol forcing are small.

Our conclusions indicate that the total anthropogenic aerosol forcing has been underestimated due to an underestimation of the IMO2020 ERF. This could be interpreted to contradict conclusions of Zelinka et al. (2026) that recent observed trends reaffirm established bounds on aerosol-cloud interactions. On closer inspection this is not necessarily the case. First Zelinka et al.’s primary aim is to constrain global ERF_{aci} and low cloud feedback predominantly with the long term trends in the observational records, with a focus on whether existing relationships from cloud controlling factors (including sulphate aerosol) can explain 2023 extremes. In contrast the IMO2020 signal (if detectable; Zhang et al. (2025) highlight the difficulties here) would be regionally concentrated, even within specific Ocean basins, and only emerge in the post 2020 record compared against the pre 2020 record. Zelinka et al. (2026) also find that North Pacific and Atlantic increasing SST and decreasing sulphate aerosol optical depth (SAOD) are the dominant contributors to the increasing trend in downwelling shortwave radiative anomalies induced by low-level clouds (R_{cl}d), and played a significant role in the 2023 extremes. Just as our analysis cannot separate the aerosol forcing component of the post 2020 CERES CRE anomalies from cloud feedback induced anomalies, the analysis in Zelinka et al. (2026) cannot separate what portion of either the aerosol component or the SST component is attributable to the forcing from IMO2020.

We have also added discussion of how our results fit in with the existing estimates and uncertainty ranges of the total anthropogenic aerosol ERF including constraints from Zelinka et al. (2026) on L431:

“Our results imply that shipping sulphate forcing, and hence the pre-2020 total anthropogenic aerosol forcing, have been underestimated. At the upper end of our sensitivity range (Ait_10) the implied increase is 0.41 W m^{-2} (Figure A1). Although this increase is 37% of the IPCC AR6 central estimate, it lies within the 90% confidence ranges of both that assessment, $-1.1 [-1.7 \text{ to } -0.4] \text{ W m}^{-2}$ (Forster et al., 2021), and the observation-based ERF_{aci} estimate of $-1.20 [-2.01 \text{ to } -0.59] \text{ W m}^{-2}$ reported by Zelinka et al. (2026).”

Comment 4:

“Discrepancies in Regional Constraints: The comparison with regional satellite observations in Section 4.4 requires a more critical, balanced interpretation. While the default Base model underestimates Diamond’s (2023) observed annual mean corridor forcing, the updated Ait_10 experiment yields an annual mean of 1.3 W m^{-2} , which overestimates the observation ($\sim 0.5 \text{ W m}^{-2}$) by more than a factor of two. Furthermore, the large gap between Zhang et al.’s (2025) regional analysis sub-total (0.074 W m^{-2}) and the model’s global 0.41 W m^{-2} is dismissed too easily, glossing over the risk of model over-sensitivity due to grid-scale emission dilution.”

There are two separate issues to address here.

The first is the comparison with Diamond’s (2023) (D23) shipping corridor Twomey effect forcing estimates. The criticism that this was not discussed or presented in sufficient detail is fair and we have made several significant changes to address this.

In the pre-print we were comparing the total model estimated CRE with D23’s Twomey-only MODIS estimated CRE. To improve this we have directly compared our model effective radius (R_e) response to D23’s and used the fractional R_e change (same methodology as D23) to estimate the Twomey-only cloud forcing component and compare to D23’s estimates and confidence intervals. In the original text we had quoted D23’s forcing estimates which were from the centre of the shipping lane. We have now expanded the region to D23’s full analysis region which reduces the average values and significantly reduces the risk from model grid scale emission dilution. These updated forcing results are now shown prominently in Figure 6 alongside the comparison to Zhang et al.’s (2025) regional cloud forcing estimates. The discussion in this section has also been updated around the new Figure 6, which replaces the previous table 2. The overestimation in Ait_10’s annual mean Twomey forcing compared to D23 is also discussed in more detail. We have also added a new Appendix A Figure A1 which shows all of our experiments against the SEA shipping corridor R_e and Twomey effect forcing constraints from D23. Figure A1 shows that Ait_High_4.5 is within or below D23’s confidence intervals for the IMO2020 perturbation in both the annual mean and SON, and has better agreement with D23’s estimates than Base.

Secondly, we must clarify that the gap between Zhang et al.’s (2025) (ZEA25) 0.074 W m^{-2} regional sub-total and our global model ERF of 0.41 W m^{-2} is not related to grid-scale emission dilution. Zhang’s three analysis regions are roughly $3000 \times 3000 \text{ km}$ and our model grid is approximately 150 km in the subtopics any resolution driven biases would show up in this comparison. The difference between ZEA25’s 0.074 W m^{-2} and our 0.41 W m^{-2} is mainly that our result is a global average, while ZEA25’s estimate assumes that the forcing from IMO2020 outside of

their 3 rectangular analysis regions is zero. In addition, our ERF includes the direct effect and cloud radiative effect whereas ZEA25's is purely shortwave CRE. Therefore, the two estimates are not directly comparable.

This point is already made in the text (lines 284-288):

“Zhang et al. (2025) also report a global forcing estimate of 0.074 W m^{-2} , but this figure is calculated by assuming zero forcing outside their three analysis regions, which together cover only ~4.2 % of the global surface area. It therefore represents the contribution of these three regions to the global IMO2020 ERF rather than a global constraint, and is consistent with a substantially larger total forcing once contributions from other areas are included.”

In order to make this clearer we have calculated the equivalent global subtotals using the exact same analysis regions and methodology as Zhang et al. (2025). The results are included alongside the absolute regional means, which were already in the text in Table 2 in our new Figure 6 (panel c).

The discussion around this in section 4.5 has also been updated accordingly. As well as updating our D23 analysis regions to better account for GCM emission dilution effects, the discussion of resolution effects has also been updated to clarify this (Sect. 4.5.2, L408-413):

“It has been shown that GCM resolution can overestimate ship-induced SW cloud forcing through the loss of sub-grid Nd variability (Possner et al., 2016), which is important for the corridor-scale comparison with D23. However, any such resolution-driven overestimation would be evident over the large ZEA25 analysis regions, where our upper forcing values underestimate the satellite-derived forcing by a factor of two. Close agreement between our model output and the estimates from D23 and ZEA25 should not necessarily be expected, given that they are derived in different ways and under different underlying assumptions. This is highlighted by the disagreement between the D23 and ZEA25 estimates relative to our results”

Responses to “Specific Comments”:

Comment:

P8 L191-193: The text notes that Ait_{88nm} has a larger total ERF than Ait but a smaller ΔCRE term, attributing the difference to noise. This, and the results in Figure 3 (particularly the RA in High 4.5 and Ait) seem to be at odds with the statement that ‘IRF and clear-clean adjustments remain relatively stable across the experiments’. Please expand this discussion.

We agree that the phrase "relatively stable" was imprecise, particularly for the clear-clean RA term. We have expanded this discussion to quantify the variation in each component, to note that some variation in IRF is physically expected given the perturbations to primary sulphate mass and size, and to discuss the clear-clean RA residual term results in more detail. The revised text (Sect. 4.2 L257-267) now reads as follows.

"Figure 6 summarises the IMO2020 ERF across the experiment suite and decomposes the response into ΔCRE , IRF, and clear-clean RA following Ghan (2013). The IMO2020 ERF shows a high sensitivity to changes in shipping-emission assumptions, spanning a range of 0.10 to 0.41 W m⁻² across our experiments. The ERF changes between experiments are dominated by changes in ΔCRE (ranging from 0.07 to 0.32 W m⁻²), with IRF remaining a relatively stable positive contribution (0.05 to 0.10 W m⁻²), and clear-clean RA consistently the smallest term, with some sign variation between experiments (-0.05 to 0.02 W m⁻²).

Increases in the direct effect (IRF) are expected due to increased sulphate aerosol lifetime and total burden from increasing emission altitude and primary fraction (Ahsan et al., 2023). Changing the primary sulphate emission radius also alters the direct effect, since scattering efficiency per unit mass depends on particle size, although this impact is modest in our experiments, likely in part because the emitted particles grow rapidly from their initial size (Fig. 2). Variation in clear-clean RA across experiments likely reflects sampling uncertainty due to internal variability rather than meaningful physical differences. The mean value across experiments is a small negative residual of -0.01 W m⁻², likely representing increased surface longwave emission in response to the land surface warming from removed aerosol."

We also clarify the Ait_88nm and Ait comparison (L273-276):

"The total ERF for Ait_88nm is larger than Ait, though we attribute this to noise in the residual clear-clean RA term and the larger direct effect in Ait_88nm. There is overlap between the standard error ranges of the two experiments and the ΔCRE term is larger in the Ait experiment, following the expected sign in response to aerosol number."

P9 L195: There is a double word typo in going from from 2.5%.

Deleted duplicate.

P13 L300: Typo in text; change gobal to global.

Corrected.

RC2

The manuscript “Recent Modelling Studies Systematically Underestimate the Warming from IMO2020 Shipping Regulations” use a single climate model (UKESM1.1) to conduct sensitivity experiments to quantify the impact of sub-grid scale emission processes on the estimated Effective Radiative Forcing following the IMO regulations for SO₂ shipping emissions.

As for aerosol ERF in general, there are large uncertainties related to magnitude of the ERF due to IMO2020, and a sensitivity study like this is very useful to quantify the impact of these assumptions on the modelling results. However, the conclusion the authors draw (also displayed in the title), based on a single model study and a highly uncertain process include many other processes than the sub-grid scale emission parameterizations, may be overstated. The last sentences of the main text highlight this limitation and recommend that this should be investigated in other models.

We agree with the point that the conclusions were potentially overstated, based on the evidence originally presented. This has been addressed in our responses to the first reviewer comments, especially their comments 1 and 2. Our updates to the manuscript, in particular the new section 4.1 and revisions to section 4.5, provide more evidence in support of the default model configuration (ERF of 0.1 W m⁻²) being an underestimate. Additionally we have made several changes to the framing and wording of the paper (detailed in our responses to reviewer 1), including a change to the title and changes to the abstract and conclusions to better reflect the remaining uncertainty. Other processes which could contribute to the model size distributions are also now explicitly discussed in section 4.1, along with the potential model dependence of these results.

Other comments:

L14: Can Δ CRE be observed from satellites?

You are right Δ CRE cannot be directly observed, these estimates we are referring to are derived from satellite observations using different methods to construct counterfactual fields and then use this to estimate the IMO2020 cloud forcing.

We have updated the abstract and removed the use of Δ CRE entirely to make it easier to understand and L14-15 has been updated from: “underestimates regional Δ CRE values relative to published satellite observations.” to: “We also find improved agreement with two independent satellite-derived constraints” in order to reflect this.

L21: reflecting additional solar radiation. Delete additional.

Deleted (now L27)

L29: Define SO_x. 8 Tg per year is not clear when you use the term SO_x.

This was an error we have corrected it to SO₂ (now L35)

To avoid confusion we have now introduced our definition of SO_x as a collective abbreviation for sulphate aerosol (SO₄) and sulphur dioxide (SO₂) when we first use it on L37: “cooling effect of shipping SO₂ and SO₄ (collectively SO_x) emissions”

L54: Are there any conclusions from Ahsan et al., 2023, the multi-model study for different emission choices, that can be highlighted in this paper, as this paper is only a single model study?

We have added this L287-290:

“UKESM1 was among the 4 out of 11 tested models with the strongest net radiative response to increased SO₂ emission altitude in Ahsan et al. (2023). Therefore, assuming this sample is representative of CMIP6 models, we would expect the altitude effect we find to be smaller on average. The UKESM1 cloud and radiative responses to increased f_SO₄ in Ahsan et al. (2023) are close to the multi-model medians.”

Figure 2: Again you use SO_x which is not clear. The unit in Fig 2a is Tg yr⁻¹, which does not make so much sense for a spatial plot. Use per square meter.

Agreed, the sulphur emissions are expressed as SO₂ only in the units of the plot. We have updated the figure caption to change SO_x to SO₂ to reflect this and match the plot colorbar.

Fig. 2a has been updated and now shows emission flux density [kg m⁻² s⁻¹].

L170: Under the header observational data only CERES data are given. The cloud radiative effect derived from the CERES data is not only dependent on the aerosols, but also feedback processes. Would be useful with more discussions related to this, as this is the only observational-based product directly used. Figure 5. CERES CRE anomaly is presented as observed. These are not directly observed and assumptions are made. How robust are the CRE anomalies from CERES? You also mention that the estimate in Hansen et al. has been scrutinized for potentially conflating aerosol forcing with coupled SST–cloud feedbacks. In your simulations, you use fixed SSTs, but compare the results with CERES data where SST has evolved. How valuable is this comparison?

We have added explicit acknowledgement of this limitation in our methodology used here and included a discussion of the low cloud feedback component in the CERES CRE anomalies (see response to reviewer 1's comment 3 for more details). Additionally, this section now includes more prominent analysis of published MODIS-derived regional IMO2020 cloud forcing constraints in Figure 6. The methodology of analysing the modelled IMO2020 CRE pattern correlations with CERES post-2020 CRE anomalies and comparing to the expected random correlations of CRE anomalies in the pre-2020 record

has been previously used by Gettelman et al. (2024). The main value of applying it here is to demonstrate that the increases in cloud forcing from IMO2020 in our experiments occur in areas that correspond to large post 2020 CERES cloud radiative anomalies. We have clarified in the text that this comparison does not serve as a constraint for the cloud forcing from IMO2020 but it does indicate our higher estimates to be more plausible than had there been no significant spatial correlations with the CERES anomalies. This has been explicitly clarified in the revised text (Sect. 4.5.3 L415-429).

CERES CRE anomaly is presented as observed. These are not directly observed and assumptions are made. How robust are the CRE anomalies from CERES?

On the use of "observed" for CERES CRE anomalies: we retain this term, following standard usage for CERES EBAF products (e.g. Loeb et al., 2021; Gettelman et al., 2024), but we now define our use of it in the methods section 3.5 (L184-185):

"Hereafter, "observed" fluxes refer to these products, in which TOA fluxes are derived from measured radiances via angular distribution models and clear-sky fluxes are filled using coincident Moderate Resolution Imaging Spectroradiometer (MODIS) retrievals; CRE uses the total-region clear-sky product, consistent with the model diagnostics."

The published regional monthly uncertainties (2.5 W m^{-2} all-sky, 5 W m^{-2} SW clear-sky; Loeb et al., 2018) apply to the absolute radiative flux absolute values; our multi-year anomalies from a single instrument record largely cancel systematic errors and average down random ones, leaving instrument stability ($\sim 0.3 \text{ W m}^{-2}$ per decade) as the residual. The uncertainty on the CRE anomalies is therefore dominated by natural cloud variability, which the pre-2020 random-correlation baseline in our method is designed to quantify.

L187: Do you need "updated" here?

Removed.

L204: SEM not defined in the main text.

Corrected, removed abbreviation as only used once.

Figure 4: This figure was a bit confusing. Took me a while to understand that the region was the aggregate of the three listed. As Fig 4d is first introduced, move this to 4a? On line 205-210, to help the reader, refer to the different panels in Fig. 4 when explaining the results of the figures.

We have updated the figure caption (now Figure 5) to clarify that it is an aggregate of three separate regions. We have also swapped panels (a) and (d) as suggested so that Fig. 5a is first introduced, and we have added additional panel references to help guide the reader (L293-294).

L280: What kind of satellite measurements do Zhang et al use?

We now specify CERES and MODIS in the methods Sect. 3.5 and Sect. 4.5.1 (L375).

L308: SEA not defined in the main text.

SEA is defined at first use (L291 in the original manuscript now L374), and the analysis region is specified in the Figure 7 caption.

L325: the aerosols are not emitted (still a large fraction of SO₂).

Updated from: “sulphate aerosol emitted from shipping.”

To: “sulphate aerosol formed from shipping emissions.” (now L443)

Citation: <https://doi.org/10.5194/egusphere-2026-2654-RC2>

RC3

This is a short, interesting, and useful article summarizing a set of global climate modeling experiments to test the effects of assumptions regarding commercial ship sulfur emissions and their changes, on the climate response to the IMO 2020 marine commercial shipping sulfur regulations. The key finding is that existing studies reporting results on the post-IMO radiative forcing change may be underestimating the magnitude of the effective radiative forcing (ERF) change.

The paper reports on a series of model experiments systematically changing three assumptions regarding the emissions: (1) emission height; (2) fraction of sulfate emitted as primary aerosol; (3) the size of the primary sulfate aerosol emitted. The experiments conducted show that all three impacts (1-3 above) are important. True emission heights from large container ships are significantly higher than assumed in many studies (near surface, or ~20 m), and this study demonstrates that increasing emission height from 20m to 100-320m altitude increases the ERF by ~40% (this is the case for the base case and for the case where primary emissions are placed into the Aitken mode). Changing the mode of the primary emissions from the base state (50% accumulation and 50% coarse mode) to the Aitken mode increases the ERF by a greater fraction (~90% for emission into a 60nm centered Aitken mode and 300% for 44nm). Finally, increasing the fraction of sulfate emitted as primary particles from 2.5% to 4.5% increases the ERF by 60% from the base case, and has a similar fractional increase when the primary emissions are in the Aitken mode. All these effects together can increase the ERF by a factor of 3-4, with an increase in ERF of as much as 0.3-0.4 W/m² possible from IMO2020. A forcing change of this magnitude would incur significant changes in global mean temperature and so this result is of major potential importance for our quantification of radiative forcing and global temperature changes over the coming decade or more.

The manuscript is well-written, and the results are very clear. It also has an excellent summary of the previous literature documenting the global ERF associated with marine shipping emissions. Several papers in that prior literature have already made estimates that are comparable to the present study, but most of these studies were conducted prior to IMO2020, and so have not received as much attention as the post IMO studies carried out in the last 2-3 years. I think this manuscript is of interest to readers of ACP and could be published with almost no revisions, but I do have a few suggestions that the authors might wish to consider.

SPECIFIC COMMENTS: A significant question I have pertains to how ship plumes are modeled in a global model where there is no concentrated ship plume representation. In the ESM used here, emissions are directly placed into a large model gridbox. I would appreciate if the authors could include a discussion of whether these results would be different if a spreading plume model were introduced to treat the ship emissions. Ship plumes spread at about 2 km per hour, so even in two days, ship emissions barely cover the scale of a ESM model gridbox. Is this likely to change the results here? Coagulation and heterogeneous chemistry occurring on particles is likely to be nonlinear, so these effects could potentially be important. It is hard for me to believe that the results would be the same if plumes were treated in a more realistic way.

We agree that this is important and should be discussed. We have added text to Section 2 addressing this explicitly (Sect. 2, L93-101):

“We therefore test increasing f_{SO_4} and reducing primary sulphate emission size in UKESM1.1. GCMs dilute shipping emissions instantaneously into grid boxes of order 100 km, whereas real plumes disperse over hours to days (Chosson et al., 2008; McMichael et al., 2024), undergoing nonlinear in-plume coagulation, condensation, and chemistry that depend on meteorological conditions (Stevens et al., 2012; Tian et al., 2014). No globally uniform set of parameters can represent this complexity, and GCM grid-scale emission dilution in isolation likely overstates the aerosol forcing (Possner et al., 2016; Stuart et al., 2013). Accordingly, we evaluate our updated parameters against background size distributions sampled away from a shipping lane, with fresh plume intercepts removed, targeting the aged, diluted aerosol that grid-scale emissions represent (Sect. 4.1). The details of our experimental design choices are presented in Sect. 3.3, along with a discussion of the uncertainty surrounding modelled SO_x emission altitude.”

We do not claim the model resolves plume processes; the f_{SO_4} and primary size parameters are effective emission properties describing the aerosol delivered to the grid scale after unresolved in-plume processing. This is why we now constrain them against the ACRUISE background size distributions (Sect. 4.1), from which fresh plumes were removed: the nonlinear in-plume processes the reviewer highlights are implicitly included in this comparison rather than neglected.

We have also added the fact that the ERF sensitivities we identify here would likely vary if emission plumes were better represented as a caveat at the end of section 4.2:

“An explicit plume treatment would likely modify the magnitude of these sensitivities, but not their direction. The default parameters substantially underestimate aerosol number in a key shipping region (Sect. 2), so any treatment correcting this, plume-resolved or parameterised, must increase CCN-active aerosol number and therefore the forcing.”

Additionally the final sentence of the conclusions now mention shipping plume representation as a key remaining uncertainty and area for future work (L469-472):

“For these reasons, future work targeted at better constraining the IMO2020 ERF is critical. This includes work on bridging scales to provide better integrated representations of emission plume and cloud processes in GCMs, comparable sensitivity analysis in other climate models and further work constraining model responses with observational constraints on IMO2020.”

Table 1: The authors show that using the updated/corrected version of Abdul-Razzak and Ghan (Ghosh et al. 2024) does not significantly impact the results. I assume that this is because the concentrations of ship-produced particles from marine shipping are quite low compared to what would be occurring in a realistic treatment of ship plumes. In near-field ship plumes, it has been shown that Ghosh et al. (2024) is needed to correctly represent aerosol activation in commercial shipping plumes (see discussion in Tippett et al., 2025). The authors may wish to comment on this.

Yes we only find and state in the paper that updating ARG2000 with the changes from Ghosh et al. (2025) does not have a statistically significant impact on the modelled global forcing from the IMO2020 shipping emission regulations. You are correct that other work has shown these updates to be very important for the accurate representation of activation in polluted environments and this effect would likely be visible regionally in our simulations but this analysis is beyond the scope of the current paper as we are focused on the impact on the magnitude of the global ERF.

We have added “global-mean” to the following sentence to clarify this (L252):

“We also tested the impact of updating the activation scheme from ARG2000 to ARG-G25 in isolation and found it had no statistically significant impact on the global-mean IMO2020 ERF under a wide range of emission parameterisations (Table~S1, Fig.~S2).”

We have also added the following text in our methods section (L123-125) where we

introduced the updated ARG activation scheme to help clarify this and better motivate this change:

“The use of the ARG activation scheme has been shown to lead to anomalous split ship-tracks where ship-tracks form at the edges of the ship plume, but disappear in centre of the plume owing to its non-monotonic behaviour at high aerosol concentrations (Tippett et al., 2026).”

Abstract: To say that recent studies show little spread is a little misleading. The range of model estimates are from 0.03-0.15 W/m² (Figure 1), so there is still some significant spread. The authors do correctly point out that these values are significantly weaker than several studies conducted prior to 2020.

We agree the wording was potentially misleading we have removed the word ‘agree’ from this sentence in the abstract (L4), the new version reads:

"Recent Global Climate Model (GCM) studies estimate that this led to a positive Effective Radiative Forcing (ERF) of $\sim 0.10 \text{ W m}^{-2}$ "

We think that this is fair given that 0.10 W m^{-2} is the mean of all model studies since 2020 (Figure 1) and the “ $\sim$ ” symbol highlights that this is an approximate value.

Can the authors potentially also comment on why none of the observational-model hybrid studies show forcings larger than 0.14 W/m^2 (Yuan et al., 2024)?

To our knowledge, Yuan et al. (2024) is the only model-observation hybrid estimate of the IMO2020 forcing. We agree that the manuscript did not make the hybrid nature of the method explicit, and have revised Section 4.3 accordingly. The key point, which was already made, is that the observational component of their method constrains the cloud response to a given Nd change; the IMO2020 Nd difference field itself is derived entirely from model output, and therefore inherits the same sub-grid emission assumptions that we highlight. The revised text reads (L350-354):

"Yuan et al. (2024) use a model-observation hybrid approach, but their IMO2020 Nd difference field is derived purely from model output, with detailed aerosol fields reconstructed from GEOS-GOCART bulk mass data using a neural network emulator (MAMnet) trained on MAM7. Because MAM7 inherits the same primary sulphate parameters as CESM from Liu et al. (2012), any structural biases regarding aerosol number and size distributions from this are propagated directly into MAMnet and so Yuan et al.'s results."

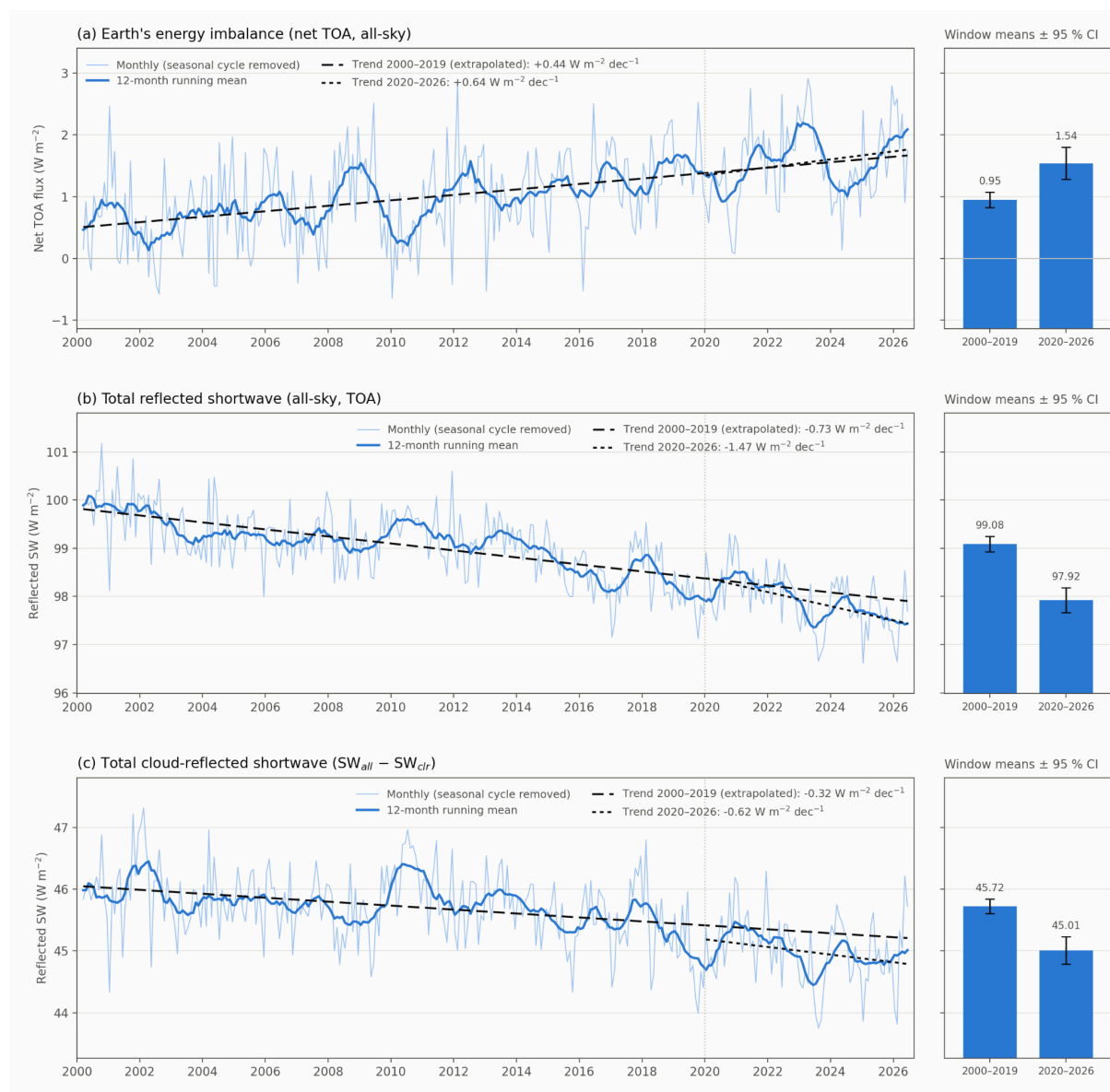
The hybrid approach therefore cannot correct an underestimate of Nd that enters upstream of the observational step, which explains why it yields a forcing in line with the purely model-based estimates.

Abstract: EEI has been increasing steadily since 2016, so it is not clear that it has surged since 2020. See e.g. Mauritsen et al. (2025). How much of the EEI change has been realized post 2020?

We did not intend to imply that EEI has surged since 2020. The sentence refers to the magnitude of the imbalance over the post-2020 period, which is the only period to which IMO2020 can contribute, and is compatible with the steady rise since 2016 noted by the reviewer; "surge" refers to global temperature, not EEI. The sentence has been revised to remove the "2 to 4 times" magnitude claim, but the reference to the post-2020 period is retained for this reason:

"A larger IMO2020 ERF would explain more of Earth's energy imbalance since 2020 and the recent global temperature surge."

Roughly one third of the EEI has been realised since 2020 $\sim 0.5 \text{ Wm}^{-2}$ / 1.5 Wm^{-2} but you are correct that this has not exceeded the expectation based on the linear trend in EEI in the pre-2020 record (see plot of CERES data below). We have added this plot to the SI (Figure S9).



Line 43: Line 43: Should probably cite Diamond et al. (2023) at this point, which provides purely observational constraints on the albedo change from IMO 2020. Diamond is cited later on.

We are specifically referring to model estimates only here (now on L47), Diamond (2023) is referenced earlier on L35 in the original manuscript (now L41).

Line 80-88: Papers from the Monterey Area Shiptrack Experiment (MAST) in 1994 also show that emitted size distributions very close to ships are mostly in the Aitken mode. For example, Hobbs et al. (2000).

Thank you for the pointer. We have added Hobbs et al. (2000) at L91 in support of the observed Aitken-mode emission size distributions.

Cloud adjustments to aerosol: Line 168-170: I would encourage the authors to explain in clearer terms the Ghan (2013) "clear-clean radiative adjustments". I don't fully understand the significance of these estimates, and what their impact is on this study. I assume it somehow accounts for the fact that CRE changes are not the same as the ERF because there are direct effects involved (changes in the clear sky radiation). I found this to be a little confusing as currently described. It also means that the reader cannot see how significant the cloud adjustments are compared with the Twomey effect. Can the adjustments simply be estimated as $ERF - IRF$? If so, the adjustments are larger than the Twomey effect, which I find confusing because under marine cloud brightening simulations, the UK ESM shows very weak cloud cover adjustments (see e.g. Rasch et al., 2024, Fig. 5). Can the authors explain why commercial shipping induces large cloud adjustments, but marine cloud brightening barely induces any? This is particularly pertinent because Diamond et al. (2020) found significant negative LWP changes in the shipping corridor associated with commercial shipping. This is directly opposite to the ESM results presented here. The authors should comment on this. It could potentially lead to a consistency with observations if the ESM here is exaggerating the cloud adjustments.

We have edited and added to the methods text (Sect. 3.4, L176-181) describing the Ghan (2013) ERF decomposition to make clearer what each term represents. To clarify here: the decomposition uses diagnostic radiation calls with aerosol removed ("clean") and with clouds removed ("clear"). The IRF is the direct effect of the aerosol on radiation (ERF_ARI); ΔCRE is the change in cloud radiative effect evaluated from the clean fluxes, and contains both the Twomey effect and cloud adjustments (ERF_ACI); and the clear-clean RA is the residual change in the aerosol-free, cloud-free flux, i.e. rapid adjustments linked to other changes in the atmospheric state and to surface albedo. The three terms sum to the ERF, so $ERF - IRF$ gives ΔCRE plus the clear-clean RA rather than the cloud adjustments. The Ghan (2013) decomposition cannot separate cloud adjustments from the Twomey effect, this point is addressed below.

It is important to clarify that we do not claim in the paper that the cloud adjustment component of the UKESM1.1 modelled IMO2020 ERF is large and none of our arguments in the paper require it to be. The point that we make at the end of section 4.4 (L364-367) is that studies which report the IMO2020 radiative forcing as the sum of only the Twomey and direct effect are missing what will be an additional likely positive (though uncertain) term from the net cloud adjustment response to aerosol. Additionally, we show that cloud adjustments increase with the cloud forcing change in our experiments (as a result of increasing N_d change) (Figure 5c,f).

We agree that the corridor LWP response is an important concern and we have added some analysis (Fig. S7) and discussion to section 4.5.2 in order to address this alongside our expanded comparison to Diamond's (2023) satellite-based estimates in Figure 6 and Figure A1 (see response to RC1 comment 4 for more

details on the additional analysis).

Additional discussion in section 4.5.2 (L396-413):

“In most experiments the SEA shipping corridor IMO2020 LWP change is small (Fig. S7), with confidence intervals spanning zero. This is consistent with UKESM's near-neutral LWP response to the large-scale Holuhraun SO₂ perturbation, which is well within the observational constraint (Malavelle et al., 2017; Wang et al., 2026). Two of the Aitken-mode experiments (Ait and Ait_10), however, produce robustly negative corridor LWP changes in response to IMO2020. This is opposite in sign to the increase implied by the LWP reduction from shipping aerosol found by Diamond et al. (2020). This suggests that the modelled total SEA corridor cloud brightness forcing could overestimate the real-world response at the correct Twomey effect magnitude. However, the correct sign of the large-scale LWP response to shipping aerosol is not well constrained. This is why we focus on the Twomey-only comparison with D23, and the total large-scale CRE comparison with ZEA25, which includes the net effect of cloud adjustments. On the other hand, UKESM1 substantially underestimates the cloud cover response to sulphate aerosol (Wang et al., 2026). This is one candidate explanation for the gap between all our model experiments and the larger satellite-derived totals in ZEA25.”