

Reply to the reviewers' comments on:

“Systematic Observation-Based Estimate of Effective Radiative Forcing from Aerosol–Cloud Interactions”

We would like to thank the reviewers for their constructive and thoughtful comments, which have helped improve this paper. Below please find a point-by-point reply to all of the reviewers' comments (in blue).

In addition to revising the manuscript according to the reviewers' comments, we also identified and corrected a minor calculation inconsistency in the estimated inversion strength (EIS) cloud-controlling factor. While MERRA-2 and CMIP6 used the exact formulation, CAMS and JRA-3Q originally relied on an alternative approximation (Ceppi et al., 2024). We updated the calculation across CAMS and JRA-3Q to match the exact formulation. This harmonization slightly adjusts some observational figures and values, but does not affect any of the study's qualitative or quantitative conclusions.

Reviewer #1:

Overall, this is an important contribution to the literature on this topic of ERF_{aci} which is still hotly debated. I appreciate the effort to reconcile various recent estimates of ERF_{aci}, or at least document which analytic decisions lead to the major differences. While it is unsatisfying that the uncertainty range of the final report estimate is still just as large as in AR6, it is useful to know how smaller or larger estimates of the forcing have emerged. I have two major comments.

Reply: We would like to thank the reviewer for acknowledging the importance of this work, and for providing these constructive comments. Below we addressed all of the reviewer's comments. Specifically, we added additional important discussions and caveats to explain and expand on points raised by the reviewer.

1. Regarding the use or not of the activation function (dN_d/dA): Is this not more physical? You show that it performs worse against the RFMIP “ground truth”. But do you have a physical interpretation for why?

Reply: We thank the reviewer for this comment. In principle, we agree that explicitly resolving the aerosol-to-cloud pathway through droplet activation (N_d) is conceptually more physical than relating an aerosol proxy directly to the cloud radiative effect, since it makes the microphysical mechanism explicit rather than leaving it implicit in a single empirical regression coefficient. However, greater physical realism in statistical models does not necessarily translate to superior predictive skill. In fact, in this case, our results show that within the statistical framework used here, this two-step “activation” route consistently underperforms the direct proxy route in terms of its regression predictive ability, suggesting that this physical realism improvement weakens the model predictive skill for reasons speculated below.

We first note an important clarification regarding the “standard” (direct proxy) method. As implemented here and in Wall et al. (2022), this approach regresses cloud radiative effect

anomalies directly against an aerosol proxy (sulfate mass concentration), without any a priori assumption about how the proxy relates to droplet number concentration (N_d). The proxy serves purely as an empirical predictor whose regression coefficient is fit directly against the radiative target; no one-to-one (or any specific) relationship between sulfate and N_d is assumed. We would characterize this differently from the description in Park et al. (2025), whose framing suggests the standard method implicitly assumes a one-to-one aerosol-to- N_d conversion; we do not believe this assumption is inherent to the direct proxy regression as applied here.

Turning to why the activation route performs worse: we believe the weaker (less negative) global ERF_{aci} obtained via the activation route, is at least partly attributable to regression dilution. When a predictor is measured with substantial random error (as is the case for satellite-retrieved N_d , which carries well-documented retrieval uncertainty (Grosvenor et al., 2018)), the resulting regression coefficient is systematically attenuated toward zero. This biases the inferred aerosol sensitivity, and hence the derived ERF_{aci} , toward a weaker magnitude.

However, this explanation does not carry over in the same way to our CMIP6 perfect-model evaluation. There, N_d is diagnosed directly from model output rather than retrieved from satellite observations, removing the classical measurement-error source of dilution, yet the activation route still underperforms the direct proxy route in this idealized, noise-free setting (Fig. 2a, in the main text). We therefore suggest two further, non-mutually-exclusive explanations specific to this result. First, the two-step mediation compounds statistical uncertainty: the activation-route ERF_{aci} estimate is the product of two independently fitted regression coefficients (aerosol-to- N_d and N_d -to-CRE) rather than one, allowing errors from each step to propagate into the final estimate. Second, even error-free N_d may not fully mediate the total aerosol-to-cloud-radiation pathway in every model: the column-maximum N_d metric used here (following Park et al., 2025) is a single, vertically-unresolved quantity that may not capture all aerosol-driven cloud adjustments contributing to the radiative response in a given model. Consistent with this, the errors in the activation route vary in direction across individual CMIP6 models rather than showing the systematic, one-signed bias that dilution would predict, supporting compounding statistical uncertainty over dilution as the likelier explanation for the CMIP6 result specifically.

We have added this reasoning to the Discussion (Sect. 4), as follows.

"Although explicitly resolving the aerosol-to-droplet pathway is conceptually more physical than an empirical direct-proxy regression, greater physical realism does not guarantee better predictive skill: the activation route underperforms the standard route consistently, in both the CMIP6 evaluation and the opportunistic experiments. We suggest this partly reflects regression dilution: N_d serves as the predictor in the second regression step, so retrieval noise in satellite-derived N_d (Grosvenor et al., 2018) attenuates its coefficient toward zero. This does not fully explain the CMIP6 result, however, where N_d is diagnosed directly from model output rather than retrieved, yet the activation route still underperforms (Fig. 2a). There, we attribute the degradation instead to compounding uncertainty from fitting two sequential regressions and to column-maximum N_d imperfectly capturing the full aerosol-to-cloud pathway in every model, consistent with prediction errors varying in sign rather than following a systematic bias."

A very relevant reference for your discussion on aerosol proxy:

Hailing Jia et al., Optimal choice of proxy for cloud condensation nuclei reduces uncertainty in aerosol-cloud-climate forcing. *Sci. Adv.* 12, eaea4828(2026)..DOI:10.1126/sciadv.aea4828

Reply: Thank you for this very relevant reference, which we are now citing in Sect. 4: *“In addition, these results are consistent with an independent assessment by Jia et al. (2026), who also find AI and surface sulfate to be a lower-bias proxy than aerosol optical depth.”*

2. Does ERF_{aci_low} dominante? Is the ERF_{aci} from non-low clouds really negligible? A lot rests on this one sentence caveat in L401.

These references all suggest that ACI in cold clouds can be substantial.

Zelinka, M. D., T. Andrews, P. M. Forster, and K. E. Taylor (2014), Quantifying components of aerosol-cloud-radiation interactions in climate models, *J. Geophys. Res. Atmos.*, 119, 7599–7615, doi:10.1002/2014JD021710.

Alexandri, F., Müller, F., Choudhury, G., Achtert, P., Seelig, T., and Tesche, M.: A cloud-by-cloud approach for studying aerosol–cloud interaction in satellite observations, *Atmos. Meas. Tech.*, 17, 1739–1757, <https://doi.org/10.5194/amt-17-1739-2024>, 2024.

Duran, B. M., N. J. Lutsko, and C. J. Wall, 2026: Aerosol–Ice–Cloud Interactions in a Perturbed Parameter Ensemble. *J. Climate*, 39, 4183–4203, <https://doi.org/10.1175/JCLI-D-25-0724.1>.

In particular, Duran et al. 2026 quantifies this ERF_{aci_ice} = -0.43 W/m². This is about 50% of the ERF_{aci_low} you quantify. If you add these together it substantially increases the ERF_{aci} (more negative) and the ECS (more positive). Your results currently do not really deviate from AR6 at all, but if there is a nonnegligible ERF_{aci_cold} that may not be true at all.

The authors should at the very least substantially expand this one-sentence caveat. Preferably you would provide a more detailed quantification of either why you now believe this to be small or what the total low+non-low ERF_{aci} actually is.

Reply: We thank the reviewer for these references and for raising this important point. We have expanded the caveat in Sect. 5 accordingly (quoted below); the core of our response concerns the scope of comparison our study is designed to make.

The previous CCF-based studies we compare our results against (Wall et al., 2022; Park et al., 2025; Zelinka et al., 2026) all compute ERF_{aci} from low-cloud radiative anomalies only, yet each compares its resulting estimate directly against assessments of the total ERF_{aci} (IPCC AR6, WCRP), implicitly treating the low-cloud contribution as an estimate of the total forcing. This line of thinking follows Bellouin et al. (2020). Since our goal is to systematically evaluate and reconcile the spread among exactly these studies, we adopt the same low-cloud-only scope and the same comparison target.

We have modified the following text accordingly:

"Finally, following the CCF-based studies we evaluate (Wall et al., 2022; Park et al., 2025; Zelinka et al., 2026), our framework computes ERF_{aci} from low-cloud radiative anomalies only and compares this directly against total- ERF_{aci} assessments (IPCC AR6, WCRP); we retain this same scope in order to evaluate and reconcile these methods on the terms they adopted. However, it is important to note that aerosol-cloud interactions in ice and mixed-phase clouds are not necessarily negligible (Zelinka et al., 2014; Duran et al., 2026). Hence, explicitly incorporating the non-low-cloud pathway remains an important direction for future work."

We also added an analysis to test our methods on low-cloud only ERF_{aci} , in the perfect model evaluation, the analysis details and figures are in our reply to Reviewer #2's 3rd comment.

Minor comments:

L66: There is also considerable uncertainty on the total pre-industrial to present-day aerosol changes. This is worth emphasizing.

Reply: Thank you for your comment, we agree and have added a caveat noting this (Sect. 5, new sixth point):

"Sixth, considerable uncertainty remains in the total pre-industrial to present-day change in anthropogenic aerosol emissions and burdens (Carslaw et al., 2013; Bellouin et al., 2020), which our Monte Carlo procedure captures only through CMIP6 inter-model spread in $\Delta\log_{10}(AI)$ and may not fully represent."

L149: Is the surface albedo calculated with clear-sky surface fluxes or TOA fluxes? This was unclear.

Reply: Thank you for pointing this out. Surface albedo is calculated using surface (not TOA) clear-sky fluxes. This matches the standard implementation of Zelinka et al. (2012) kernel method, since the shortwave kernel's sensitivity depends on the reflectivity of the surface beneath the cloud. We have revised the text at L149 to make it clear:

"Because the shortwave kernel depends on the underlying surface albedo, we calculate a monthly climatological surface albedo for each grid cell as the ratio of surface upwelling to downwelling clear-sky shortwave radiation"

L309: Over Eastern North America, I would argue that MERRA and CAMS capture these trends at significantly different magnitudes. Why?

Reply: We agree the original wording understated the discrepancy and have removed "slightly" from the text, which now reads:

"The MERRA-2 and CAMS reanalyses capture these trends at different magnitudes (Figs. 3a and 3c), providing an additional test of sensitivity to reanalysis choice."

MERRA-2 and CAMS are independent reanalysis products, each with its own data assimilation system and underlying set of assimilated observations, and some divergence between them in regional aerosol trends is therefore expected. Isolating the specific cause of the magnitude difference over Eastern North America would require dedicated work that is beyond the scope of this study. This is precisely why our framework treats reanalysis choice as an explicit, quantified source of structural uncertainty (Fig. 1b) rather than relying on a single reanalysis product, and why we report results from all six reanalysis combinations throughout (Figs. 3, S6) rather than selecting one as ground truth.

Fig 4b: Why is the largest ΔAI over the Arabian Peninsula?

Reply: AI is calculated from total column aerosol optical depth across all species, whereas SO_4 and SAOD are specific to sulfate. The large $\Delta \log_{10}(AI)$ signal over the Arabian Peninsula, which is not mirrored in the sulfate-based proxies, is therefore likely driven by non-sulfate aerosol changes in this region. In addition $\Delta \log_{10}(AI)$ reflects the ratio of present-day to pre-industrial AI rather than absolute magnitude. Thus, this large $\Delta \log_{10}(AI)$ signal over the Arabian Peninsula might be simply because the AI was very small there during pre-industrial times.

Reviewer #2:

General Comments

This manuscript presents a systematic framework for evaluating methodological choices in CCF-based estimates of ERFaci. The authors examine the sensitivity of ERFaci estimates to the choice of aerosol proxy, spatial domain, regression framework, reanalysis dataset, and treatment of aerosol activation. This systematic comparison is highly valuable and timely, particularly given the substantial spread among recent observation-based estimates of ERFaci. With revisions that fully address the points below, I believe this work could make a valuable contribution to the literature on observational constraints of aerosol–cloud radiative forcing.

Reply: We would like to thank the reviewer again for these constructive comments. Below, we address all of the reviewer's points. In particular, following the reviewer's comments, we have extended our observational opportunistic experiments analysis to evaluate the activation pathway, added a like-for-like low-cloud evaluation in CMIP6, clarified the physical justification and trade-off for including land, and detailed how our uncertainty propagation explicitly accounts for calibration ensemble size.

Major Comments

1. One of my main concerns is the use of a very limited number of models to derive the model-based bias correction. In the analysis, the key relationship appears to be based on only six models. While the Monte Carlo procedure propagates uncertainty in the fitted relationship, it does not overcome the fundamental limitation introduced by the small sample size and the struc. With these few models, a single model can exert substantial leverage on the regression slope and can strongly inflate or reduce the estimated correlation. The resulting relationship may therefore be highly sensitive to individual models and may not represent a robust emergent relationship across the models. Particularly, a Pearson correlation of $r=0.51$ based on six models is not statistically distinguishable from zero at conventional significance levels. An emergent-constraint-style model-based bias correction requires a statistically robust relationship between the CCF-predicted and ground-truth ERFaci estimates. In the absence of such evidence, I do not think this relationship provides a sufficiently strong basis for quantitatively adjusting the primary observation-based ERFaci estimate. A public comment by Chanyoung Park supports this concern. There it is shown that one of the major conclusions of the paper is changed dramatically if a larger set of models is sampled. Specifically, when a larger set of models is considered, the explicit inclusion of activation rates improves the perfect-model cross-validation, rather than degrades it as the manuscript currently asserts. This is a serious limitation of the current manuscript and one that must be addressed prior to publication.

Reply: We thank the reviewer for this important comment, and for engaging with the public comment from Chanyoung Park, which provides a genuinely useful quantitative test of this concern. We agree the small model sample is a real limitation of our work, and we'd like to address the reviewer's points in three parts: (1) why our CMIP6 evaluation ensemble is smaller

than that used in the analysis presented in the public comment, and what we gain from that choice; (2) a new, fully independent test of the activation pathway using the opportunistic experiments, which was not part of the original manuscript; and (3) how the small sample size is reflected in our uncertainty quantification.

1. Model sample size versus consistency with the observational pipeline

We agree that a larger CMIP6 sample is, all else being equal, statistically preferable. However, expanding the sample size in this context is not "all else equal": it requires relaxing the correspondence between the model evaluation pipeline and the observational pipeline it is meant to validate. There are two defensible ways to conduct this evaluation, and they trade off against each other.

Our observational ERF_{aci} is derived by taking CERES-retrieved cloud fractions and fluxes binned by cloud-top pressure (CTP) and optical depth. Using this dataset we compute the non-obscured low-cloud radiative effect at every grid cell via Eq. (1) from the main text, using a CTP threshold of 680 hPa. This allows us to separate low from non-low cloud contributions even in scenes where both are present. To evaluate this specific pipeline in the CMIP6 dataset, we require models that provide ISCCP-simulator output. Only the simulator enables identical joint CTP-optical depth binning and the non-obscured low-cloud decomposition (Eq. 3 from the main text) across all grid cells, including those with overlapping cloud layers. This restricts us to the 11 models with full ISCCP-simulator output (6 of which additionally provide the AI proxy needed for our top-performing configuration), but it means the model test isolates the same physical quantity, over the same grid cells, using the same masking treatment, as the observational estimate it is meant to validate.

Our understanding of the approach used in Park et al. (2025) and in the public comment's analysis (and we would welcome correction if we have mischaracterized it) is that the perfect-model $ERF_{aci,true}$ is instead obtained by classifying each grid cell into a single cloud regime following Webb et al. (2006), combined with a radiative-kernel-based decomposition of all-sky versus clear-sky TOA fluxes. Using this approach, the corresponding model evaluation is restricted to the subset of grid cells classified as low-cloud-dominated, rather than the full domain. This approach does not attempt to reproduce the specific non-obscured, mixed-cloud-aware decomposition used in the observational CERES pipeline. That is presumably how this method can be extended to a larger model sample (10-12 models in the public comment) - **but at the cost of no longer testing the same quantity, over the same grid cells, that the observational estimate actually relies on. In other words, we face a clear trade-off: performing a direct, apples-to-apples comparison across a smaller model ensemble, or relying on an uncertain model-observation comparison across a larger sample. We opt for the former.**

We expanded our discussion of this trade-off in the manuscript, after the CMIP6 evaluation results:

“This differs from Park et al. (2025)’s own perfect-model comparison, which instead classifies grid cells into cloud regimes following Webb et al. (2006) and thereby supports a larger model sample. We prioritize consistency with our observational pipeline, restricting our comparison to the 10 models providing the ISCCP-simulator output it requires; the same ranking of methods holds when using a low-cloud-only ground truth or an ocean-only domain instead (Figs. S8, S9 in the Supplement).”

2. An independent test: extending the opportunistic experiments to the activation pathway

Because we recognize that a 10-model CMIP6 comparison alone cannot fully settle the activation pathway versus direct regression concern, we have added a second, fully independent test of the activation pathway that does not depend on CMIP6 model sampling at all. We extended our opportunistic-experiment framework (Sect. 3.2, Fig. 3) to explicitly include the activation route, using

$$\Delta R_{SW} = \sum_{CCF_{met}} \left(\frac{\partial R'_{SW}}{\partial CCF'_{met}} + \frac{\partial R'_{SW}}{\partial \log_{10} N'_d} \times \frac{\partial \log_{10} N'_d}{\partial CCF'_{met}} \right) \times CCF_{met} + \frac{\partial R'_{SW}}{\partial \log_{10} N'_d} \times \frac{\partial \log_{10} N'_d}{\partial proxy'} \times \Delta proxy,$$

where ΔR_{SW} is the change in low-clouds short-wave radiation, CCF_{met} is one of the 6 meteorological CCFs, $\log_{10} N_d$ is the log of droplet concentration and proxy is the aerosol proxy CCF, and prime indicates an anomaly of a variable. Specifically, we use it to predict the observed long-term shortwave low-cloud radiative trends over Eastern North America and Eastern Asia, as we did previously for the standard route. This test uses real, observed aerosol and meteorological trends rather than a model ensemble, so it is not subject to the small-sample leverage concern the reviewer raises for the CMIP6 comparison. The revised Fig. 3 (now showing both the standard, solid markers, and activation, hollow markers, routes) and the revised Fig. S6 (extending this comparison across all 11 regression configurations, for both regions), shows that the activation route consistently predicts the observed trend less accurately than the standard route in most regression configurations, and especially in the OLS 1×1 configuration that is the focus of this discussion.

We think this is the most important point in response to the reviewer's concern. We now have two evaluation frameworks that are as independent from one another as we can construct: a perfect-model CMIP6 comparison (limited to 6 models, but matched as closely as possible to the observational pipeline) and a real-world opportunistic-experiment comparison. Both independently point to the same conclusion - that the activation route underperforms the standard route, particularly for OLS 1×1. We do not think this agreement is guaranteed, and we believe it substantially strengthens our confidence that the CMIP6 result is not simply an artifact resulted from the small model-sample, even though we cannot rule out that a larger, differently-constructed model sample (as in the public comment) would shift the CMIP6 comparison specifically. We revised the manuscript to present both lines of evidence, and expanded the first caveat, as follows:

Methods: *"For the activation route, the meteorological CCF sensitivities combine their direct effect on the low-cloud radiative anomaly with their indirect effect mediated through N_d , while the aerosol-proxy sensitivity is obtained solely through the N_d -mediated pathway."*

Results (opportunistic experiments): *"Consistent with the CMIP6 evaluation, the activation route also underperforms the standard route across all proxies in the OLS 1×1 configuration."*

Fig. 3 caption: *"The activation route underperforms the standard route in the OLS 1×1 configurations, in all proxies."*

Conclusions (First caveat): *"expanding the model ensemble in future phases of CMIP would further refine these constraints, and test the robustness of the methodological rankings identified here."*

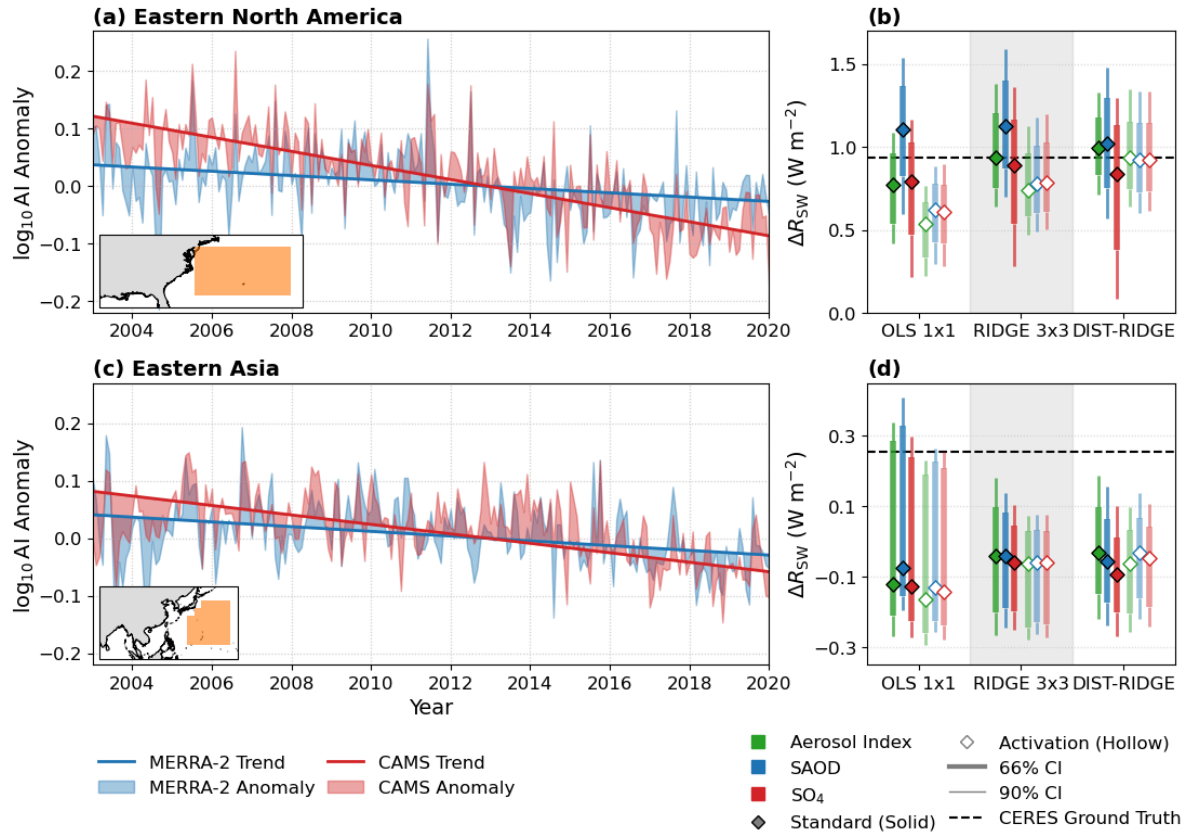


Figure 3. Validation via opportunistic experiments in regions with well-characterized recent aerosol trends. **(a, c)** Time series (2003–2020) of $\log_{10}(\text{AI})$ anomalies and linear trends from MERRA-2 and CAMS over Eastern North America and Eastern Asia, respectively, shown for AI as the primary proxy identified in the CMIP6 evaluation. Both regions experienced sustained multi-year trends in aerosol loading driven by documented changes in sulfur emissions, providing a strong observational test of method skill. **(b, d)** Predicted decadal change in shortwave low-cloud radiative effect (ΔR_{SW} , defined as the 2012–2019 average minus the 2003–2010 average) from the three top-performing regression configurations of the CMIP6 evaluation, compared against the CERES observed change (dashed black line). Predictions are shown for both standard and activation routes, with all the reanalyses combinations predictions combined in a single distribution, for each one of the proxies. Error bars show 66% and 90% confidence intervals derived from a stationary bootstrap (Sect. S2 in the Supplement). The AI OLS 1×1 includes the ground truth ΔR_{SW} in its 66% CI in both regions, supporting its selection as the primary methodology. The activation route underperforms the standard route in the OLS 1×1 configurations, in all proxies. Results for all configurations are presented in Figure S6 in the Supplement.

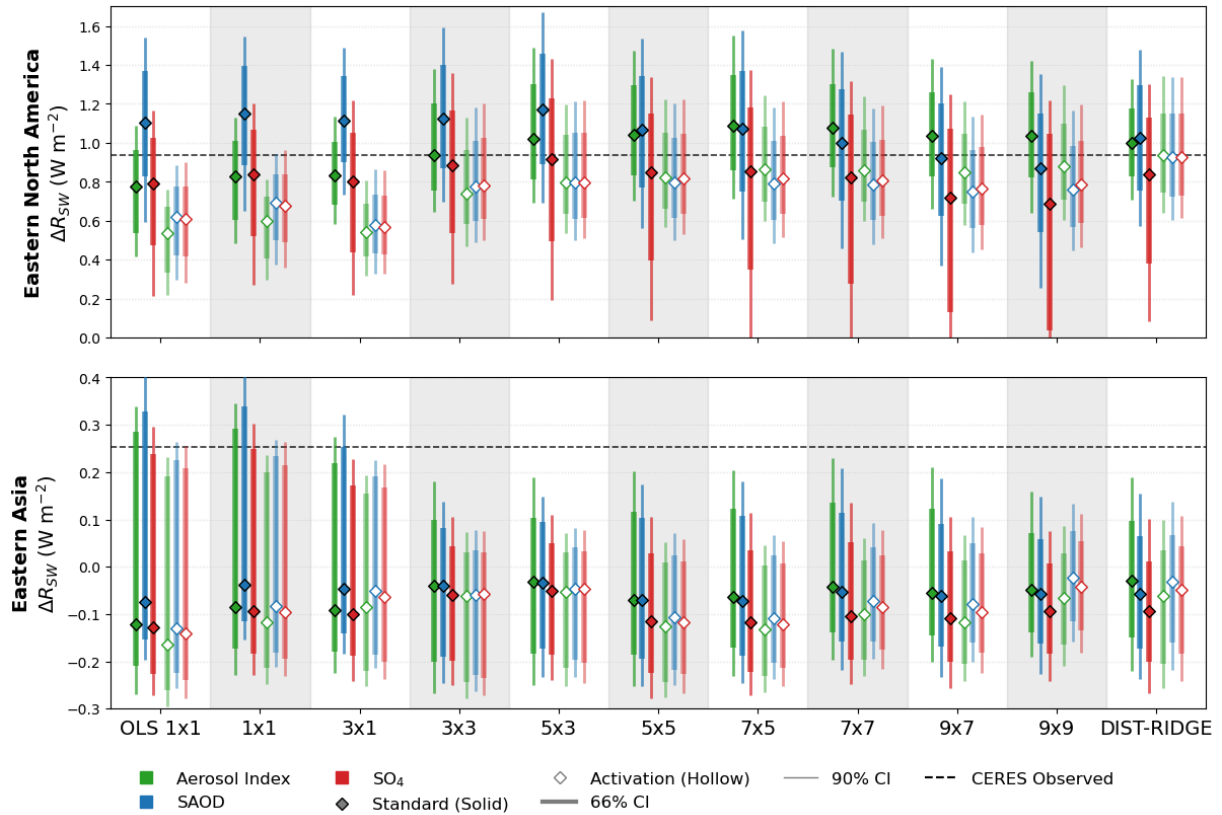


Figure S6. Full opportunistic experiment validation results for all tested regression configurations, aerosol proxies, and reanalysis combinations, extending Fig.3 of the main text. Rows correspond to the two validation regions (Eastern North America, top; Eastern Asia, bottom). Within each panel, the x-axis organizes configurations by regression method and domain size, progressing from local OLS through increasing Ridge domain sizes up to the global DIST-RIDGE, and the y-axis shows the predicted change in shortwave low-cloud radiative effect (ΔR_{SW} , $W m^{-2}$), defined as the 2012–2019 average minus the 2003–2010 average. Each entry is the combined estimate across all reanalysis combinations for each of the aerosol proxies (AI, SAOD, SO_4). Vertical lines show the 66% (thick) and 90% (thin) confidence intervals derived from a stationary bootstrap (Sect. S2 in the Supplement). The dashed horizontal black line indicates the CERES-observed ΔR_{SW} ground truth. Predictions are shown for both standard and activation (hollow markers) routes.

3. Bias-correction and uncertainty estimates using small sample size

On the specific point about Fig. S7: the reviewer is correct that a Pearson $r = 0.51$ from 6 models is not distinguishable from zero at conventional significance. More broadly, we agree that a correlation coefficient is not a meaningful diagnostic at this sample size - with only 6 points, r is highly sensitive to individual models and its associated significance test has little power. We have therefore removed the Pearson r annotations from the relevant SI figures (Figs. S2-S5, S7) and rely on root mean square error (RMSE) alone as our metric for comparing

configurations throughout the evaluation. RMSE measures predictive accuracy directly rather than resting on a significance framework this sample size cannot support.

This is also how the bias correction itself is used: it does not treat the fitted slope as a single, known number. The weakness of the fit is instead captured by the 90% prediction interval in Fig. S7, and it is this interval (not just the central slope and intercept) that is propagated through the Monte Carlo procedure. The correction shifts our central estimate only modestly, from an uncalibrated -0.69 Wm^{-2} (90% CI: -1.05 to -0.39 Wm^{-2}) to a bias-corrected -0.85 Wm^{-2} (90% CI: -1.59 to -0.19 Wm^{-2}). The calibration step roughly doubles the width of the confidence interval. This widening is the mechanism by which our limited ability to calibrate against a six-model ensemble is honestly reflected in the final reported uncertainty.

We have considered the alternative of using the regression-coefficient-based uncertainty method from Wall et al. (2022)/Park et al. (2025)/Zelinka et al. (2026) (as we do for the exploratory comparison in Fig. 1) instead of the emergent-constraint bias correction for our final estimate. We chose not to, because that method has no mechanism for capturing the mismatch between the raw CCF-predicted ERF_{aci} and the true forcing that Fig. S7 reveals - it would report a falsely narrower uncertainty, regardless of how well or poorly the CCF regression actually recovers the true forcing. We would rather report a wider, more honest uncertainty that explicitly reflects our limited ability to calibrate and evaluate against a small model ensemble, than a narrower one that omits this source of error altogether. We expanded the relevant paragraphs in the manuscript to include this reasoning and the uncertainty inserted by the small model ensemble.

In the Methods, we expanded the bias-correction paragraph with: *“We adopt this approach, propagating the fit's prediction interval through the Monte Carlo procedure, rather than the regression-coefficient-based uncertainty used in previous studies (Wall et al. (2022), Park et al. (2025), Zelinka et al. (2026)) , because only the former captures the uncertainty of the CCF analysis performance in predicting the ERF_{aci} itself.”*

In the Results, we expanded the bias-correction paragraph with: *“This correction shifts our central estimate modestly, from an uncalibrated -0.69 Wm^{-2} to a bias-corrected -0.85 Wm^{-2} , while roughly doubling the width of the resulting 90% confidence interval; this widening, which reflects the limited size of the calibration ensemble, is a dominant contributor to our final reported uncertainty.”*

In the caption of figure S7, we added the CI change between the calibrated and uncalibrated estimates: *“For the observations predicted ERF_{aci} , the calibration shifts the central estimate from an uncalibrated -0.69 Wm^{-2} (90% CI: -1.05 to -0.39 Wm^{-2}) to a bias-corrected -0.85 Wm^{-2} (90% CI: -1.59 to -0.19 Wm^{-2}); the roughly twofold widening of the interval reflects the additional uncertainty introduced by calibrating against a small, 6-model ensemble.”*

2. The manuscript adopts the combined ocean-and-land domain as the primary analysis domain, mainly because including land reduces the uncertainty in the domain-to-global scaling derived from the aerosol-only experiments. However, it is not clear that this statistical advantage in the model-derived scaling is directly transferable to the CCF-based analysis. The CCF framework has primarily been developed and applied to

marine low clouds, for which the selected predictors have well-established physical relationships with cloud variability over ocean. Over land, boundary-layer structure and surface heterogeneity can differ substantially from those over the ocean. The authors should therefore provide a clearer physical justification for extending the CCF framework over land. Moreover, Figure S5 appears to show better predictive skill for the ocean-only domain, which is more physically consistent with the original CCF framework.

Reply: We thank the reviewer for raising this point, and agree it deserves a clearer justification than the manuscript previously provided.

Although the CCF framework was originally developed for marine low clouds, applying it over land is already established in the literature: recent low-cloud feedback studies have used the same CCF approach over combined ocean-and-land domains (Ceppi and Nowack, 2021; Ceppi et al., 2024, 2026), and we follow the same procedure and the same set of CCFs over land as those studies.

We also agree with the reviewer's underlying physical point: boundary-layer structure and surface heterogeneity over land are less well captured by our predictor set than over ocean, and this does show up as modestly lower raw predictive skill in Fig. S5 - RMSE for the three top AI configurations is 0.41, 0.34, and 0.41 Wm^{-2} over ocean-only, versus 0.45, 0.44, and 0.45 Wm^{-2} including land, a difference of 0.03-0.10 Wm^{-2} across all three cases. However, the quantity that ultimately matters for our final ERF_{aci} estimate is not the domain-level RMSE in isolation, but the total propagated uncertainty in the global estimate, which also depends on the domain-to-global scaling factor γ . Including land grid cells narrows the 95% CI half-width of γ by more than a factor of six (from 1.08 ± 0.38 to 0.87 ± 0.06), because CMIP6 models agree far more closely on this scaling when land is included than when the domain is restricted to ocean. This reduction more than compensates for the small degradation in raw predictive skill: as shown in Fig. 2b, the total uncertainty for all three top-performing AI configurations is lower for the combined ocean-and-land domain than for ocean-only. So this is a genuine trade-off between two competing sources of uncertainty: marginally lower domain-level prediction skill over land, against substantially better inter-model agreement on the scaling factor. By our own tests, the latter wins out by a small but consistent margin across all three configurations. This is also consistent with our response above on the bias-correction uncertainty: because our final uncertainty comes from how well predictions actually match the CMIP6 ground truth, rather than from regression-coefficient uncertainty alone, the reduced accuracy on land is not lost from the accounting - it is what makes this trade-off assessable in the first place. Coefficient-based uncertainty alone, as used in prior studies, would not capture as well the degraded land-cell accuracy, understating the true cost of including land.

We have revised the manuscript to explicitly state this trade-off, rather than leaving it implicit. The revised manuscript now reads:

"Although the CCF framework was originally developed for marine low clouds, extending it to land is well established in recent low-cloud feedback studies (Ceppi and Nowack, 2021; Ceppi et al., 2024, 2026). Continental boundary layers exhibit greater surface heterogeneity and complex thermodynamic structures that our meteorological CCF set captures less completely

than over ocean, which is reflected in a modest increase in raw domain-level RMSE ($0.44\text{--}0.45\text{ Wm}^{-2}$ across the top configurations over ocean and land compared to $0.34\text{--}0.41\text{ Wm}^{-2}$ over ocean-only; Fig. S5 in the Supplement). However, including land grid cells substantially reduces uncertainty in the domain-to-global scaling parameter γ , from 1.08 ± 0.38 under the ocean-only domain to 0.87 ± 0.06 under the combined domain (mean $\pm$ 95% confidence range), a more than sixfold reduction in the uncertainty range. This reflects far greater inter-model agreement on γ when land cells are included, which outweighs this small degradation in raw prediction skill, and propagates directly into a consistently lower $ERF_{aci,g}$ uncertainty across all top-performing configurations (Fig. 2b). We therefore evaluate all subsequent results over the combined ocean-and-land domain."

3. As also noted by the first reviewer, the CCF-based estimate in this study is restricted to low clouds. I understand that this choice follows the original development and primary application of the CCF framework, which focuses on low-cloud radiative variability. Nevertheless, aerosol interactions with non-low clouds also contribute to the total ERF_{aci} . More importantly, I'm also not sure about the consistency of the perfect-model evaluation. My understanding is that the RFMIP aerosol-only ground-truth ERF_{aci} includes the radiative response of all cloud types, whereas the CCF-based prediction is derived only from low-cloud radiative anomalies. If so, the comparison is not strictly like-for-like?

Reply: We thank the reviewer for pressing on this point specifically.

The reviewer's characterization is correct: in Fig. 2a (and Figs. S2-S5), the RFMIP "ground truth" $ERF_{aci,d}$ against which we compute RMSE reflects the model's full aerosol-only radiative response, restricted to the analysis domain but not restricted to low clouds, while our CCF-predicted $ERF_{aci,d}$ is derived only from low-cloud radiative anomalies. As explained in our reply to Reviewer #1's second comment, we compare against the total-forcing target deliberately, since our final uncertainty and bias correction (Fig. S7) is calibrated against each model's total RFMIP forcing in order to produce an estimate comparable to total- ERF_{aci} assessments such as AR6 and WCRP, rather than a low-cloud-only number. However, that choice does not by itself establish whether the relative ranking of configurations in Fig. 2a is sensitive to this all-cloud-vs-low-cloud mismatch in the ground truth, and this is an additional analysis we did in response to the reviewer's comment.

Using the same ISCCP-simulator-based decomposition (Eq. 3, in the main text) that isolates the low-cloud-only radiative response in our CMIP6 historical evaluation, we applied this decomposition to the RFMIP aerosol-only simulations themselves, giving a genuinely low-cloud-only $ERF_{aci,d}$ ground truth. This lets us re-run the Fig. 2a comparison fully like-for-like: low-cloud CCF prediction against low-cloud RFMIP truth, using the same set of models as the original comparison. We run it using both ocean cells only and including land cells.

The attached figure shows the result. For including land cells, the overall pattern from Fig. 2a is preserved: the standard route outperforms the activation route in the majority of configurations, and the three configurations we identified as best-performing in the original comparison (AI OLS 1×1 , AI RIDGE 3×3 , and AI Distance-Weighted Ridge) remain among the best-performing

configurations here as well, with RMSE around 0.25-0.35 Wm⁻², well below the SO₄, SAOD, and the activation configurations. For ocean cells only, we see AI with activation get similar, slightly lower, predictive skill as the standard route AI. The best configurations are still standard routes.

We take this as evidence that our methodological conclusions, both the standard-vs-activation comparison and the selection of the top AI configurations, are not an artifact of comparing against a total-forcing target, but hold under a genuinely like-for-like, low-cloud-only comparison as well, using the same model sample (all models available with the ISCCP simulator and at least one aerosol proxy variable).

We added this figure to the Supplement, and referenced it alongside the caveat we are adding in response to Reviewer #1's second comment on the low-cloud scope of our framework, noting that we tested the low-cloud-only ground truth as a robustness check and found the same qualitative ranking of methods. We retained the total-forcing RFMIP target for the final bias correction (Fig. S7) itself, since that is the quantity we ultimately need to correct against - but this additional analysis gives us confidence that doing so does not distort which configuration we select as best.

The revised caveat in the conclusions now reads: *"Finally, following the CCF-based studies we evaluate (Wall et al., 2022; Park et al., 2025; Zelinka et al., 2026), our framework computes ERF_{aci} from low-cloud radiative anomalies only and compares this directly against total- ERF_{aci} assessments (WCRP 2019 (Bellouin et al., 2020) and IPCC AR6 (Forster et al., 2021)); we retain this same scope in order to evaluate and reconcile these methods on the terms they adopted. However, it is important to note that aerosol-cloud interactions in ice and mixed-phase clouds are not necessarily negligible (Zelinka et al., 2014; Duran et al., 2026; Dagan, 2026). Hence, explicitly incorporating the non-low-cloud pathway remains an important direction for future work."*

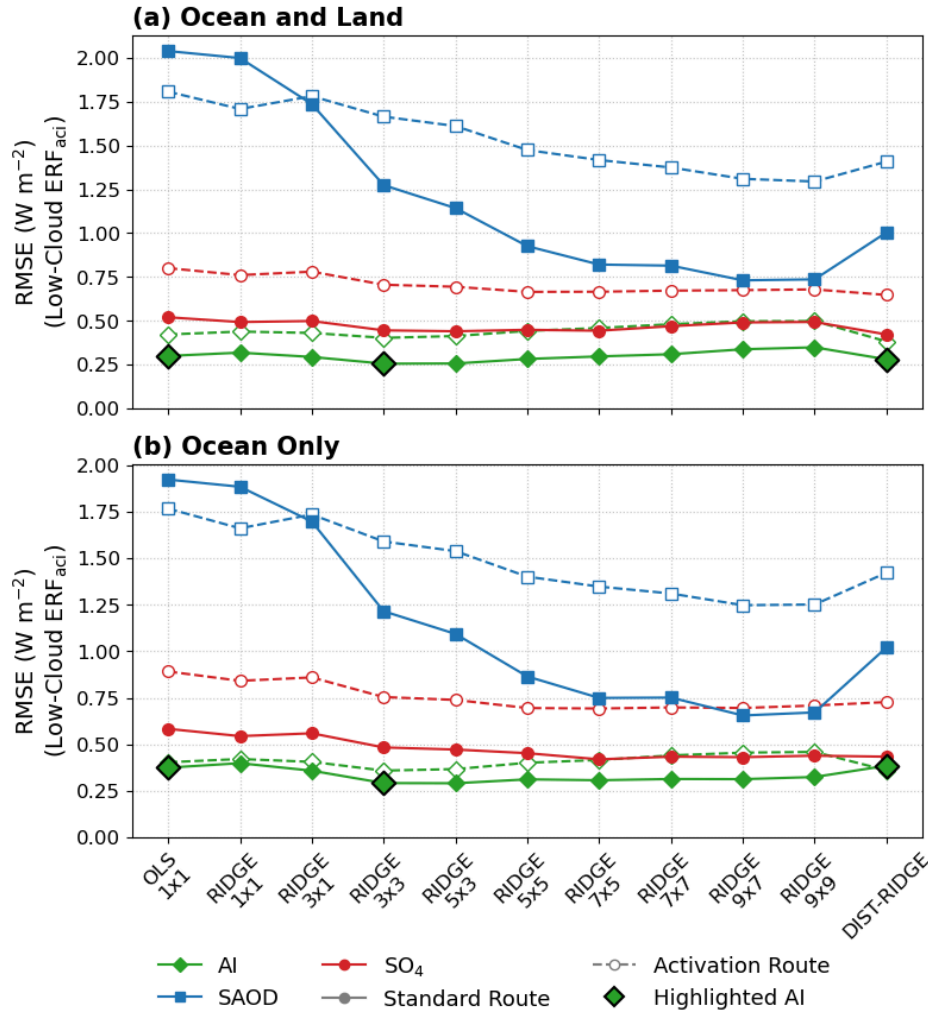


Figure S8. Evaluation of regression methodologies against a low-cloud-only RFMIP “ground truth”, as a robustness check on the total-forcing comparison in Fig. 2a of the main text. RMSE of the predicted $\text{ERF}_{\text{aci,d}}$ across all tested configurations of aerosol proxy (AI, SAOD, SO₄), regression strategy, and physical pathway, evaluated against the low-cloud-only aerosol-only forcing in the same CMIP6 models as Fig. 2a, isolated by applying the same ISCCP-simulator-based decomposition (Eq. 3 in the main text) to the RFMIP aerosol-only simulations. **(a)** Combined ocean-and-land domain. **(b)** Ocean-only domain. Solid lines denote the standard direct proxy route; dashed lines denote the intermediate droplet activation route. Highlighted markers (black outline) indicate the three best-performing AI configurations identified in Fig. 2a of the main text (OLS 1×1 , RIDGE 3×3 , and DIST-RIDGE). The regression CCF grid cell domain size around the target cell is written in each regression technique, except the Distance-Weighted Ridge that takes CCFs from the whole globe. In both domains, the activation route does not outperform the standard route across all proxies and regression strategies, and the same three AI configurations remain among the best-performing, confirming the conclusions of Fig. 2a are not an artifact of comparing against a total-forcing ground truth.

4. The perfect-model comparison appears to lead to the opposite conclusion from Park et al. (2025), with the activation pathway performing worse and therefore being excluded from the final estimate. Park et al. (2025) includes a larger model sample in its comparison, which may provide a more statistically robust assessment of the activation pathway. Because this comparison determines whether the activation step is retained, the discrepancy should be discussed further.

Reply: We thank the reviewer for this comment. In our reply to the comment above, we discussed two things relevant here: first, the trade-off between how closely the evaluation pipeline matches the observational pipeline and the resulting model sample size, which explains why our CMIP6 ensemble is smaller than the one used in the public comment's analysis; and second, a new independent test using the opportunistic experiments, which does not depend on CMIP6 sampling at all and reaches the same conclusion as our CMIP6 comparison. We'd like to add two further points here, specific to Park et al. (2025)'s own published perfect-model comparison, which we believe are at least as relevant to the discrepancy as sample size alone.

First, Park et al. (2025) do not rerun their full CCF regression (including the full set of meteorological predictors) on the CMIP6 model output. As described in their Appendix A7.3, their model-based $ERF_{aci,est}$ is instead calculated from a simplified regression that relates the low-cloud CRE (or N_d) directly to the aerosol proxy alone, without the meteorological CCFs (their Eqs. A6-A7). This raw model estimate is then rescaled by a single global factor derived from the ratio between their CCF-based and non-CCF-based $ERF_{aci,obs}$ in the observational analysis. The exact CCF regression, which is the actual methodological choice being evaluated, is not what is tested against the models; a simplified, non-CCF relationship is tested instead, and a single scalar correction is applied to approximate what the CCF-based estimate would have been. Our CMIP6 evaluation, by contrast, reruns our complete regression pipeline (Sect. 3.1) individually on each model's own aerosol, cloud, and meteorological output, exactly mirroring the procedure applied to observations. We believe this is a substantially more direct test of the specific methodological choices this study evaluates, independent of sample size.

The second point concerns the "ground truth" Park et al. (2025) use to expand their own model sample beyond the models with available RFMIP aerosol-only runs. They use the Soden and Chung (2017, SC17) method to construct a pseudo-truth $ERF_{aci,true}$ surrogate for this larger sample, justified by their reported agreement between $ERF_{aci,true}$ and $ERF_{aci,SC17}$ (RMSE = 0.14 Wm^{-2} in their Fig. 3b). We repeated this same comparison (RFMIP aerosol-only "true" ERF_{aci} versus SC17-estimated ERF_{aci}) with all grid cells (not just low-clouds dominant classified ones), across all CMIP5 and CMIP6 models for which the necessary output is available (attached figure). We find a RMSE of 0.305 Wm^{-2} for total ERF_{aci} , which is smaller, but in similar magnitude of the CCF-analysis evaluation on models RMSE. That is why we believe using this as ground truth could make the evaluation less reliable, and preferred the smaller model size.

Combined with the sample-size/pipeline-accuracy trade-off and the opportunistic-experiment evidence discussed in our reply above, we believe these two points give us confidence that our evaluation is a more faithful test of the specific pipeline this study is designed to evaluate.

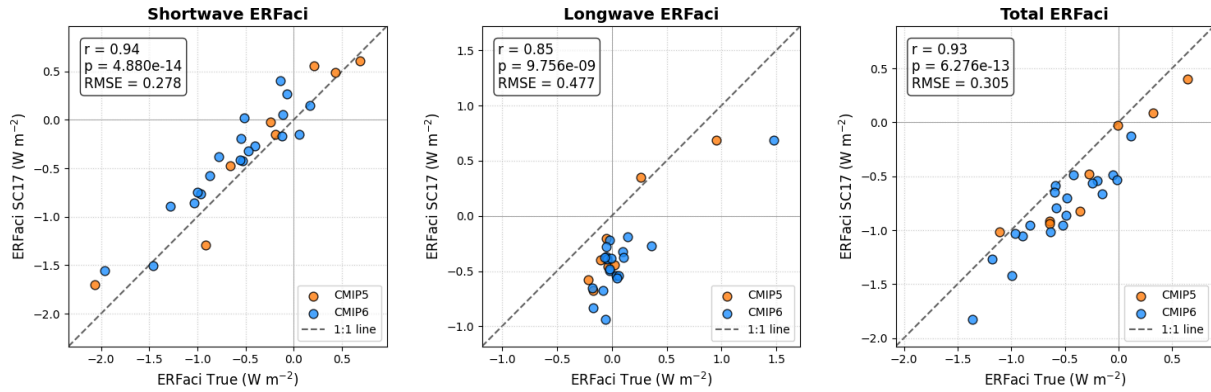


Figure R1. Comparison of global ERF_{aci} estimated via the Soden and Chung (2017; SC17) method against the aerosol-only simulation ground truth across available CMIP5 (orange) and CMIP6 (blue) models. Panels show shortwave (left), longwave (middle), and total (right) forcing components, with the 1:1 line shown dashed. For total ERF_{aci}, the discrepancy between SC17 and the ground truth yields an $RMSE = 0.305 \text{ Wm}^{-2}$, which is comparable in magnitude to the differences between the CCF regression configurations being evaluated. This demonstrates that using SC17 as an expanded pseudo-truth introduces non-negligible surrogate error.

Minor Comments

- Figure 1a: I'm concerned about including land regions in the activation-based ERF_{aci} estimates. Satellite retrievals of Nd are generally much less reliable over land because of heterogeneous and often highly reflective surface backgrounds (Grosvenor et al., 2018). Please justify this choice and clarify how retrieval uncertainty over land is treated.

Reply: We thank the reviewer for raising this. Figure 1a is intended purely as an illustrative, exploratory figure: it replicates the exact combinations of methodological choices (aerosol proxy, spatial domain, reanalysis, and physical pathway) used across the prior CCF-based studies we discuss (Wall et al., 2022; Park et al., 2025; Zelinka et al., 2026), using their own uncertainty quantification approach (Sect. S1), in order to show how sensitive the resulting ERF_{aci} is to these choices in aggregate. Beyond generating Fig. 1b's structural uncertainty budget, none of the individual estimates shown in Fig. 1a (including the land-inclusive activation-route points the reviewer highlights) are carried forward into any subsequent step of our analysis. The activation route over land+ocean is not evaluated further, corrected, or used in our final estimate; our CMIP6 and opportunistic-experiment evaluations independently rule out the activation pathway entirely (Sect. 4), so the land-versus-ocean question no longer matters for that pathway once it has been excluded.

That said, we agree the point about N_d retrieval reliability over land is worth flagging explicitly, since the uncertainty bars shown for the activation-route, land-inclusive points in Fig. 1a are derived only from the regression-coefficient-based method (Sect. S1) and do not include any additional term for the retrieval uncertainty of N_d itself. We have added the following caveat to Text S1: Central Estimates and Uncertainty for Figure 1, in the Supplement:

"MODIS-retrieved N_d carries substantial additional retrieval uncertainty over land relative to ocean, owing to heterogeneous and often highly reflective land surfaces (Grosvenor et al., 2018); this source of uncertainty is not reflected in the confidence intervals shown for the land-inclusive activation-route estimates in Fig. 1a of the main text."

2. Figure 1b: The relatively high RMSE of the activation-based estimates may partly reflect an inconsistency in the data sources used along the causal pathway? N_d is obtained from satellite observations, whereas the aerosol proxies and meteorological predictors are taken from reanalysis products. This may favor the standard route and make the activation route appear less skillful for reasons that are not solely related to the physical validity of the activation framework. Why is MODIS-derived AI not used for the analysis? Because MODIS AI is a more directly observation-based product than reanalysis-derived AI, including it could help assess whether data-source consistency affects the comparison.

Reply: We thank the reviewer for this comment. First we want to clarify a point about Fig. 1b, since we believe there may be some confusion about what it shows. Fig. 1b does not report RMSE; it shows the mean conditional standard deviation introduced into the global ERF_{aci} estimate by switching a single methodological choice (e.g., standard vs. activation route) while holding all other choices fixed, computed purely from the spread across our six reanalysis combinations in the observational analysis (Sect. S1). It is not derived from any comparison against a ground truth, and so it cannot by itself indicate whether the activation route is more or less accurate - only that switching between the two routes changes the resulting ERF_{aci} by a large amount. We suspect the reviewer's concern is more relevant to the RMSE-based evaluations in Fig. 2a (CMIP6 perfect-model comparison), where the activation route does show consistently worse skill against a ground truth, and we address that concern below.

The reviewer raises an important point: in our observational pipeline, N_d comes from MODIS satellite retrievals while the aerosol proxies and meteorological predictors come from reanalysis products, so any degradation in the activation route's performance there could in principle reflect this data-source inconsistency rather than a genuine shortcoming of the activation pathway itself. However, this concern does not apply to our CMIP6 perfect-model evaluation (Fig. 2a), which we consider the primary basis for excluding the activation route from our final estimate. There, N_d , the aerosol proxies, and the meteorological predictors are all diagnosed directly and consistently from the same model's own output, with no mixing of satellite and reanalysis sources at all. Despite this fully self-consistent setting, the activation route still underperforms the standard route (Fig. 2a). We already discuss a related point in our reply to Reviewer #1's first comment, where we note that regression dilution from noisy satellite-retrieved N_d is a plausible explanation for the weaker activation-route ERF_{aci} in observational studies specifically, but that this explanation does not carry over to the CMIP6 evaluation, since model-diagnosed

N_d has no such retrieval noise, and the activation route underperforms there too. We think this is the key evidence that data-source inconsistency alone cannot explain our result, though we agree it may still contribute to the degradation seen in the observationally-based opportunistic experiments.

On MODIS-derived AI specifically: we chose not to use it because our AI proxy is deliberately calculated to match the CMIP6 AI formulation as closely as possible (Ångström exponent from the same wavelength pairs applied to total aerosol optical depth; Sect. 2.1-2.2), so that our CMIP6 perfect-model evaluation is testing a consistently-defined quantity between the observational and model pipelines. MODIS AI is calculated from different wavelength bands than either our reanalysis-derived AI or the CMIP6 AI, so substituting it would introduce a mismatch between the observationally-derived proxy and the model-diagnosed proxy used to validate it, undermining the like-for-like comparison the CMIP6 evaluation depends on.

3. L120: Please clarify which N_d retrieval is used in the activation analysis following Gryspeerdt et al. (2022). Specifically, is the dataset based on Aqua, Terra, or a combined Aqua and Terra product? Please also state which retrieval and filtering approach is applied for the analysis.

Reply: We thank the reviewer for asking for this clarification. We follow the same choice as Park et al. (2025): the BR17 sampling strategy at the 3.7 μm retrieval wavelength, from the Gryspeerdt et al. (2022) MODIS N_d data product, combining Aqua and Terra observations. We revised the manuscript to explicitly state that:

“Finally, for the intermediate activation route analysis (Park et al., 2025), we use cloud droplet number concentration (N_d) retrieved from the Moderate Resolution Imaging Spectroradiometer (MODIS; Gryspeerdt et al. 2019a). Following Park et al. (2025), we obtain N_d from the Gryspeerdt et al. (2022) data product, which combines observations from both the Aqua and Terra satellites. Specifically, we select retrievals derived from the 3.7 μm channel filtered according to the sampling criteria of Bennartz and Rausch (2017) (termed the “BR17” sampling strategy in Gryspeerdt et al. 2022). Because the N_d dataset concludes in 2020, any analysis using N_d is restricted to the 2003–2019 period, with anomalies calculated relative to the climatology and trends of that specific timeframe.”

4. L303: It would strengthen the validation if the authors also considered episodic aerosol perturbations, as in Wall et al. (2022), in addition to the regional decadal changes examined here.

Reply: We thank the reviewer for this suggestion. We performed this test, following Wall et al. (2022) as closely as possible: we use the same region and the same two Kilauea eruption events (2008 and 2018) as an out-of-sample test, training the regression on the remaining years of data. Because our record extends three years beyond the period used in Wall et al. (2022), this is not an exact one-to-one replication of their test, but the region and events tested are identical.

The two attached figures summarize our findings. Figure R2 shows the raw time series of the CERES low-cloud shortwave anomaly alongside the three aerosol proxies (SAOD, AI, and SO4)

from both MERRA-2 and CAMS, with the eruption periods highlighted. Figure R3 shows the predicted versus observed ΔR_{SW} response to each eruption across all regression configurations, for both the standard and activation routes.

A few things stand out. First, while all three aerosol proxies show a clear, large-magnitude spike during both eruptions, the corresponding CERES low-cloud SW anomaly does not show a correspondingly large or consistent response, particularly for the 2008 eruption, where the observed anomaly is not large in absolute magnitude at all (Fig. R2). Second, of all the proxy/reanalysis combinations, only SO_4 from MERRA-2 (The proxy that was used in Wall et al. 2022) reproduces the observed anomaly reasonably well; SO_4 from CAMS does not (Fig. R3). This is despite the two reanalyses' SO_4 time series looking broadly similar to one another over the eruption periods in the anomalies timeseries, so we do not have a clear explanation for why one succeeds and the other does not. AI and SAOD do not perform well in either reanalysis. We suspect this may partly reflect that these two proxies also respond to total column aerosol, which can be influenced by above cloud base aerosol increase during large episodic emissions, which SO_4 concentration at 910hPa specifically excludes. However, we are not fully confident in this explanation given the sparse and noisy nature of the test.

We believe this kind of episodic test is examining a different question than the one our framework is built to evaluate. Eruptions like Kilauea produce aerosol proxy values outside the range spanned by the multi-year regional trends the regression is trained on. Predicting the response to such an event therefore requires extrapolating fitted regression coefficients well beyond the range of the training data, rather than interpolating within it. Consequently, we have less reason to trust that a locally fitted linear sensitivity remains valid that far outside its calibration range. Our opportunistic experiments (Sect. 3.2, Fig. 3), by contrast, test the regression against smooth multi-year regional trends of a magnitude much closer to the day-to-day and year-to-year variability the regression is actually fit on. Thus, we think this is a fairer test of the specific skill we are trying to validate: whether the fitted CCF sensitivities correctly capture the response to realistic sustained aerosol changes, rather than to isolated extreme events. On reflection, we suspect the success Wall et al. (2022) obtained in this kind of episodic test may be somewhat sensitive to their specific choice of proxy and reanalysis dataset, given how differently the various combinations perform in our own version of this test.

For completeness, we also checked whether the activation route performs any differently than the standard route in this episodic setting, and it does not: the activation route underperforms the standard route here as well, consistent with our CMIP6 and trend-based opportunistic-experiment results discussed elsewhere in this response.

Given these mixed and not fully interpretable results, we would prefer not to add this episodic test to the manuscript itself, but we wanted to share it with the reviewer directly and explain our reasoning for retaining the trend-based opportunistic experiments as our primary observational validation.

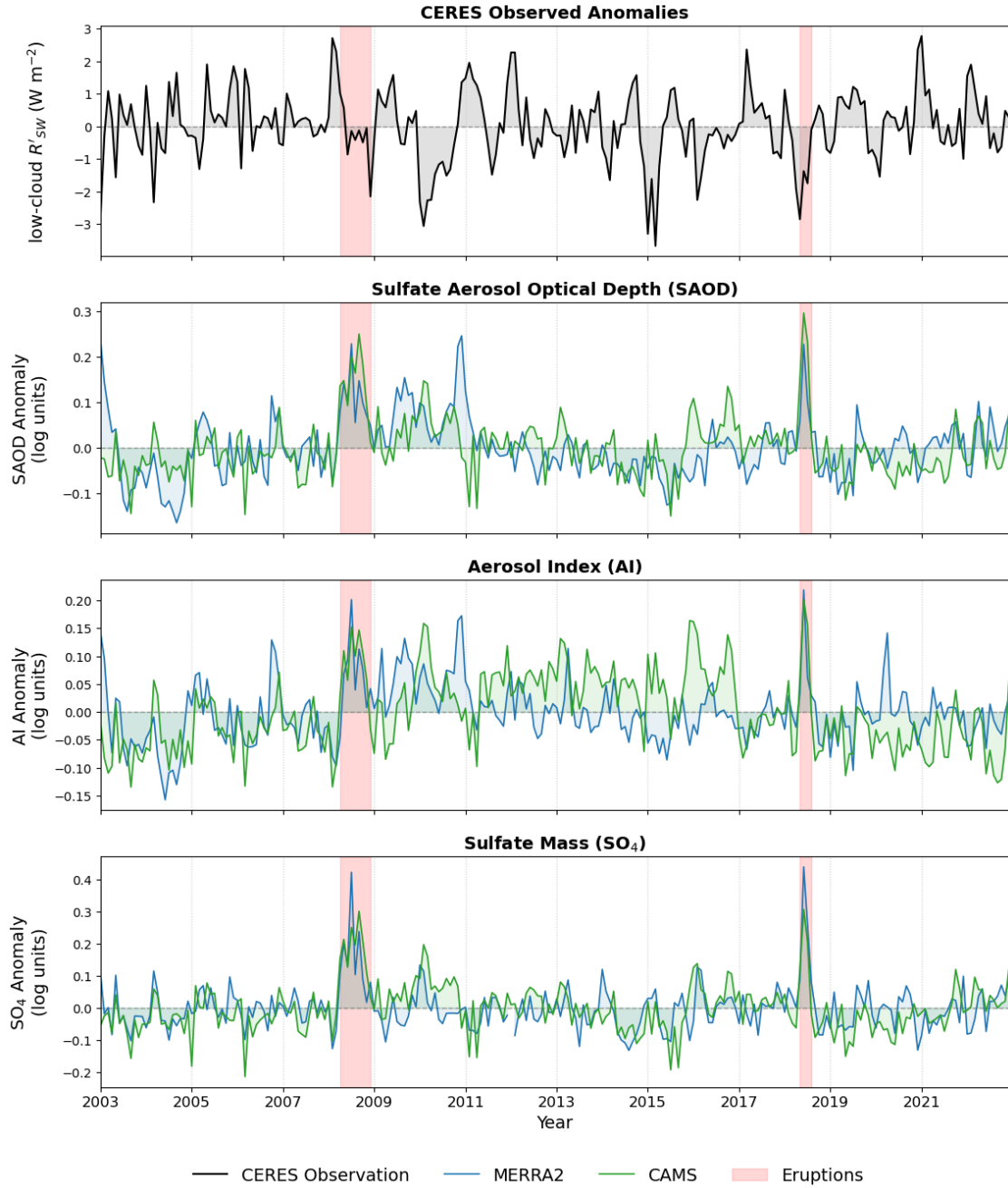


Figure R2. Monthly time series (2003-2022) of regional mean anomalies over the Kilauea domain (following Wall et al., 2022) for CERES low-cloud shortwave radiative effect (R'_{sw} ; top), and the three aerosol proxies: Sulfate Aerosol Optical Depth (SAOD), Aerosol Index (AI), and Sulfate mass concentration (SO_4) from MERRA-2 (blue) and CAMS (green). All variables are deseasonalized and detrended. Red vertical shading denotes the 2008 and 2018 Kilauea eruption periods. While both reanalyses capture pronounced, coherent spikes across all aerosol proxies during both events, the corresponding CERES low-cloud R'_{sw} does not exhibit a matching, distinct cooling anomaly - particularly during the 2008 eruption.

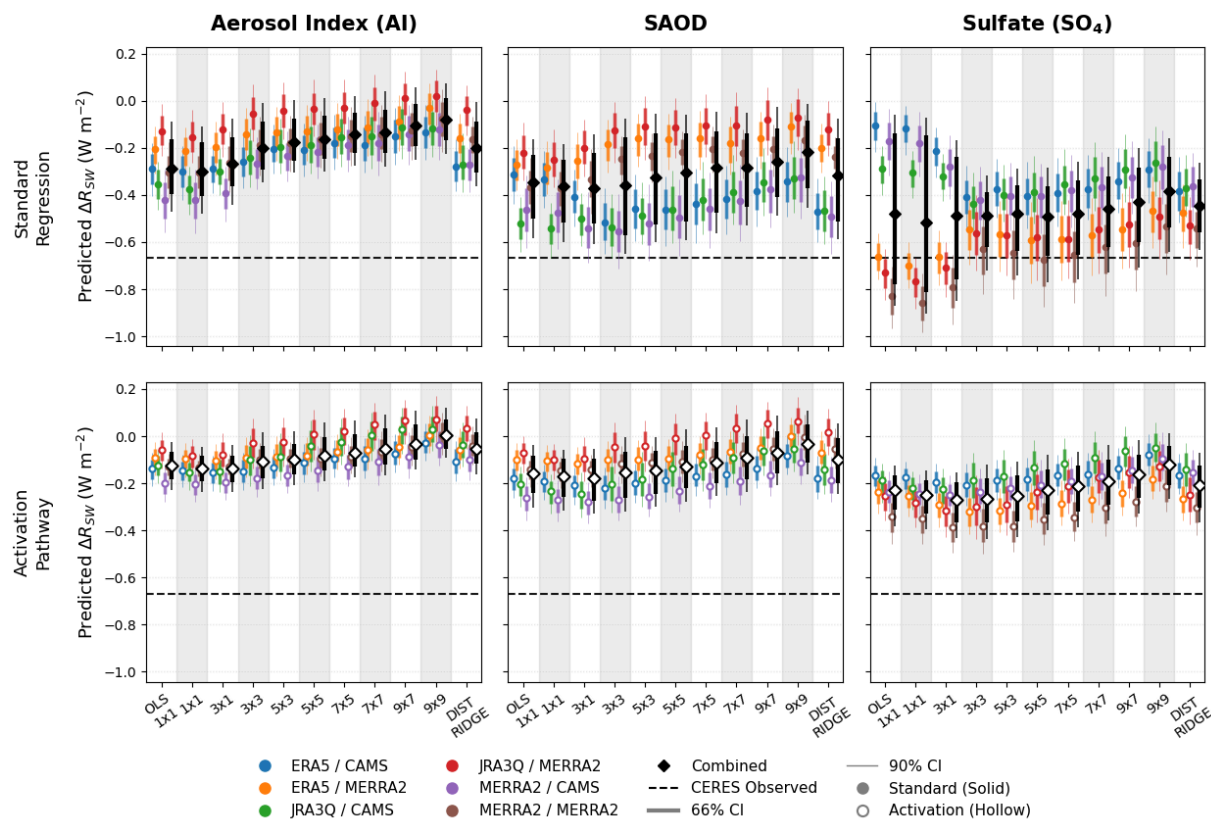


Figure R3. Out-of-sample predicted versus observed change in low-cloud shortwave radiative effect (ΔR_{SW}) for the Kilauea eruption periods across all tested regression configurations, aerosol proxies, and reanalysis combinations. The top row shows the standard direct proxy regression (solid circles); the bottom row shows the droplet activation pathway (hollow circles). Columns correspond to AI (left), SAOD (middle), and SO_4 (right). Within each panel, methodologies range from local OLS 1×1 through increasing Ridge domain sizes up to the global Distance-Weighted Ridge. Colored markers show individual reanalysis combinations; black diamonds denote the combined multi-reanalysis distribution. Thick and thin vertical lines represent 66% and 90% confidence intervals, respectively, and the horizontal dashed black line indicates the CERES observed ΔR_{SW} . Only SO_4 from MERRA-2 (the configuration used in Wall et al., 2022) reproduces the observed anomaly, while the activation pathway consistently underperforms the standard route.

5. L364: It would be helpful to clarify for readers which reanalysis dataset, or combination of reanalysis datasets, is used to derive the final observational ERFaci estimate.

Reply: This is already stated in both the Methods and Results sections.

In the Methods (Sect. 3.2), we state: "We quantify our uncertainty using Monte Carlo sampling. For each method and domain combination, we draw 1,000,000 samples. In each iteration, we randomly sample one set of aerosol proxy coefficients (from the 6 reanalysis combinations), one

CMIP6 model $\Delta Proxy$, one slope and intercept from the bias correction distribution, and one domain-to-global scaling parameter (γ)."

And in the Results (Sect. 4), when reporting the final estimate: "After propagating all sources of uncertainty, including regression coefficient uncertainty across six reanalysis combinations, inter-model spread in $\Delta \log_{10}(AI)$ across 11 CMIP6 models, calibration uncertainty from the bias correction slope and intercept, and γ scaling uncertainty, through a one-million sample Monte Carlo procedure, our best estimate of the global ERF_{aci} is -0.85 Wm^{-2} (66% confidence interval (CI): -1.23 to -0.48 Wm^{-2} ; 90% CI: -1.59 to -0.19 Wm^{-2})."

6. Figure S7 caption: I don't think the results are sufficient to support the statement that the raw CCF method systematically underestimates the magnitude of aerosol cooling. This conclusion is based on a limited number of models.

Reply: We agree with the reviewer and have removed this statement from the Fig. S7 caption.

7. Please report the corresponding p-value alongside each Pearson correlation coefficient (r) throughout the text and figures.

Reply: As discussed in our reply above regarding Fig. S7 and the model-based bias correction, we agree that a correlation coefficient is not a meaningful diagnostic at this small sample size - with as few as 6 models. With a small sample size, Pearson correlation is highly sensitive to individual models and its associated significance test has little power, so reporting it (with or without a p-value) risks conveying a false sense of statistical support. Rather than adding p-values alongside each Pearson r , we have therefore removed the Pearson r annotations throughout the manuscript and Supplement (Figs. S2-S5, S7) and rely on RMSE alone as our metric for comparing configurations. RMSE measures predictive accuracy directly without resting on a significance framework these sample sizes cannot support.

Technical corrections

1. Eq (3): Because the Zelinka et al. (2012) kernel has units of $\text{W m}^{-2} \%^{-1}$, a factor of 100 % may be required in the equation. This differs from the CERES FBCT formulation in Eq. (1), where has units of W m^{-2} .

Reply: We thank the reviewer for catching this. We have revised Eq. (3) to include the factor of 100 to correctly account for the units of the Zelinka et al. (2012) kernel ($\text{Wm}^{-2} \%^{-1}$), consistent with the fractional cloud-property changes it is multiplied against.

References

Grosvenor, D. P., Sourdeval, O., Zuidema, P., Ackerman, A., Alexandrov, M. D., Bennartz, R., Boers, R., Cairns, B., Chiu, J. C., Christensen, M., Deneke, H., Diamond, M., Feingold, G., Fridlind, A., Hünerbein, A., Knist, C., Kollias, P., Marshak, A., McCoy, D., Merk, D., Painemal, D., Rausch, J., Rosenfeld, D., Russchenberg, H., Seifert, P., Sinclair, K., Stier, P., van Diedenhoven, B., Wendisch, M., Werner, F., Wood, R., Zhang, Z., and Quaas, J.: Remote Sensing of Droplet Number Concentration in Warm Clouds: A Review of the Current State of Knowledge and Perspectives, *Reviews of Geophysics*, 56, 409–453, <https://doi.org/10.1029/2017RG000593>, 2018.

Gryspeerdt, E., McCoy, D. T., Crosbie, E., Moore, R. H., Nott, G. J., Painemal, D., Small-Griswold, J., Sorooshian, A., and Ziemba, L.: The impact of sampling strategy on the cloud droplet number concentration estimated from satellite data, *Atmospheric Measurement Techniques*, 15, 3875–3892, <https://doi.org/10.5194/amt-15-3875-2022>, 2022.

Park, C., Soden, B. J., Kramer, R. J., L’Ecuyer, T. S., and He, H.: Observational constraints suggest a smaller effective radiative forcing from aerosol–cloud interactions, *Atmospheric Chemistry and Physics*, 25, 7299–7313, <https://doi.org/10.5194/acp-25-7299-2025>, 2025.

Wall, C. J., Norris, J. R., Possner, A., McCoy, D. T., McCoy, I. L., and Lutsko, N. J.: Assessing effective radiative forcing from aerosol–cloud interactions over the global ocean, *Proc. Natl. Acad. Sci. U.S.A.*, 119, e2210481119, <https://doi.org/10.1073/pnas.2210481119>, 2022.

Zelinka, M. D., Klein, S. A., and Hartmann, D. L.: Computing and Partitioning Cloud Feedbacks Using Cloud Property Histograms. Part I: Cloud Radiative Kernels, <https://doi.org/10.1175/JCLI-D-11-00248.1>, 2012.

Community Commenter #1:

Dear Omer Roi-Cohen and co-authors,

Thank you for this interesting and timely paper. A systematic comparison of the different cloud-controlling-factor (CCF) approaches used to estimate ER_{Faci} is valuable and much needed. I was particularly interested in the perfect-model evaluation. In Park et al. (2025), accounting for the aerosol activation rate reduced the RMSE relative to estimates that did not account for activation. I was surprised that your analysis appears to reach the opposite conclusion. I have several questions and additional analyses that may help clarify the source of this difference.

Reply: We sincerely thank Chanyoung Park for his constructive engagement with our preprint and for sharing his detailed independent analyses. We greatly appreciate the time and effort invested in reproducing and discussing these comparisons with us. Below, we address each of the questions and points raised.

1) I understand that your ground-truth ER_{Faci} estimates are derived from Smith et al. (2020). Did you separate the contributions from low and non-low clouds when calculating the ground-truth ER_{Faci} ? In Park et al. (2025), we separated these contributions using the cloud classification of Webb et al. (2006). This appears important because the CCF framework assumes that aerosol–cloud interactions primarily affect low clouds, with negligible contributions from non-low clouds. Therefore, consistently restricting both the CCF-based estimate and the model-derived ground truth to low-cloud contributions may be necessary for a direct comparison.

Reply: As explained in our reply to Reviewer #2's 3rd comment above, our RFMIP "ground truth" $ER_{aci,true}$ is not separated into low and non-low cloud contributions; it reflects the model's full aerosol-only radiative response over the analysis domain. We adopt this total-response target because it matches the comparison target used by the previous CCF-based studies we evaluate (Wall et al., 2022; Zelinka et al., 2026), which likewise compare their low-cloud-derived ER_{aci} against total-forcing assessments (IPCC AR6, WCRP) rather than a low-cloud-only ground truth.

That said, we agree consistency between the CCF-based estimate and the model ground truth matters, and as described in our reply to Reviewer #2 above, we repeated the CMIP6 evaluation using the ISCCP-simulator-based decomposition to isolate a genuinely low-cloud-only RFMIP ground truth, and re-ran the comparison fully like-for-like (low-cloud CCF prediction against low-cloud RFMIP truth). The result (see the figure referenced in our reply to Reviewer #2's comment above) preserves the same overall pattern: the standard route outperforms the activation route in the majority of configurations, and the same three AI configurations remain among the best-performing even when both sides of the comparison are restricted to low clouds.

2) Have you compared the performance of the activation-rate methods separately over ocean and over the combined land+ocean domain? In the manuscript, I found this comparison only for the three best-performing methods—OLS 1x1, RIDGE 3x3, and DIST-RIDGE—but these appear to be methods in which the activation rate is not explicitly considered.

I understand that using the combined land-and-ocean domain may reduce the uncertainty in the ground-truth ERF_{aci} estimate. However, the CCF frameworks considered here were primarily developed for marine low clouds. Cloud controlling factors and aerosol–cloud relationships over land may differ because continental clouds are influenced by meteorological controls that are not fully represented in the current CCF analysis. For this reason, all three CCF studies for ERF_{aci}— Wall et al. (2022), Park et al. (2025), and Zelinka et al. (2026)—as well as the original CCF study of Scott et al. (2020), focused on ocean regions and excluded land. Therefore, comparing methods with and without activation over the combined land+ocean domain may not provide a proper evaluation of the intended marine low-cloud CCF framework. An ocean-only comparison would help determine whether the contrasting result relative to Park et al. (2025) is primarily related to the evaluation domain.

Reply: Thank you. As discussed in our reply to Reviewer #2's 2nd comment, we are aware that the CCF framework was originally developed for marine low clouds, and we agree there is some physical justification for that original scope, since boundary-layer structure and surface heterogeneity over land are less well represented by our predictor set than over ocean. However, extending CCF analysis to combined land-and-ocean domains is by now established in the literature (e.g., Ceppi and Nowack, 2021; Ceppi et al., 2024, 2026), and we follow the same procedure over land as those studies. As we detailed in that reply, including land does come with a modest cost to raw domain-level predictive skill (Fig. S5), but it more than compensates for by substantially narrowing the uncertainty in the domain-to-global scaling factor γ , so the combined domain yields lower total uncertainty (Fig. 2b). Given this, and given the precedent in the literature, we believe combined land-and-ocean is a defensible primary domain, though we agree it is not the only reasonable choice.

Following this comment, we ran ocean-only comparisons between standard and activation routes, using all configurations, with the all-cloud RFMIP ERF_{aci} as the ground truth (Figure S9). We also ran low-clouds only RFMIP ERF_{aci} as the ground truth for both ocean only and ocean+land (shown in our reply to Reviewer #2's third major comment, Figure S8). In both figures, restricting the domain to ocean-only does not change our qualitative conclusion: the activation route does not improve on the standard route. In the all-clouds RFMIP ERF_{aci} comparison attached here, the activation route's RMSE is consistently higher than the standard route, especially in the best AI configurations.

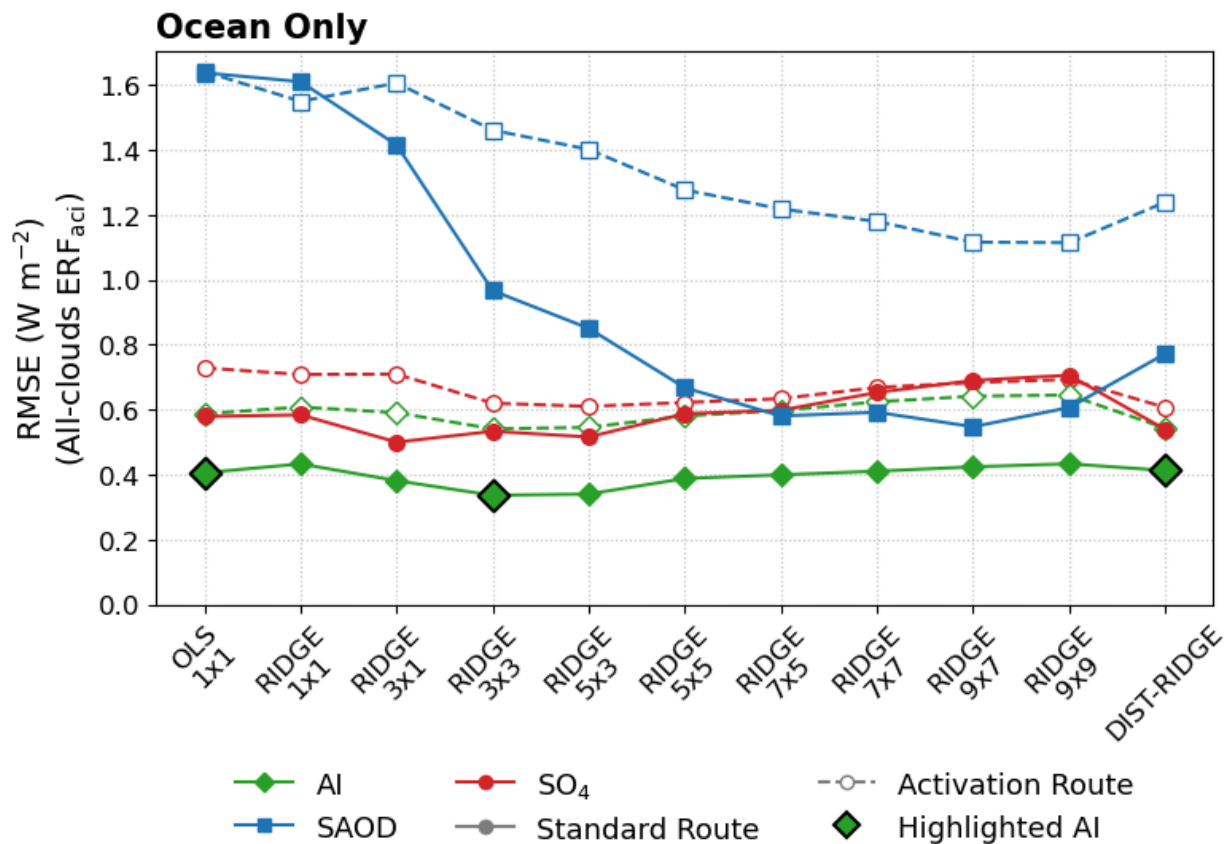


Figure S9. Evaluation of regression methodologies against CMIP6/RFMIP “ground truth” over the ocean-only domain, replicating the analysis of Fig. 2a of the main text. Root Mean Square Error (RMSE; W m^{-2}) of predicted $\text{ERF}_{\text{aci,d}}$ across all tested configurations of aerosol proxy (AI, SAOD, SO_4), regression strategy, and physical pathway, evaluated against the total (all-clouds) aerosol-only forcing in CMIP6 models restricted to ocean grid cells between 55°S and 55°N . Solid lines with filled markers denote the standard direct proxy route; dashed lines with open markers denote the intermediate droplet activation route. Highlighted markers (black outline) indicate the three best-performing AI configurations identified in Fig. 2a of the main text (OLS 1×1 , RIDGE 3×3 , and DIST-RIDGE). Restricting the domain to ocean cells preserves the primary conclusions: the standard route continues to outperform or match the activation pathway across all proxies, and the AI configurations retain the highest predictive skill (lowest RMSE).

In the low-cloud-only, ocean-only result referenced above (Fig. S8), the activation route either has the same, or lower, performance as the standard route. We didn’t find it to offer any improvement.

So while we agree with the commenter, and with Reviewer #1 on a related point in their first comment, that explicitly resolving the aerosol-to- N_o -to-CRE pathway is conceptually more physical than a direct proxy regression, our perfect-model analysis does not show this added physical realism translating into improved predictive skill, in either the combined or the ocean-only domain. We discuss in our reply to Reviewer #1’s first comment several possible

reasons for this (compounding statistical uncertainty from fitting two sequential regressions, and the imperfect nature of column-integrated N_d metrics as mediators of the full aerosol-to-cloud-radiation pathway), which we believe are more likely explanations than the evaluation domain given that the pattern persists under ocean-only as well.

We have added both figures to the SI, and referenced them in the results section:

“the same ranking of methods holds when using a low-cloud-only ground truth or an ocean-only domain instead (Figs. S8, S9 in the Supplement).”

3) I was unsure about the meaning of labels such as OLS 1x1, RIDGE 3x3, 7x5 and 9x7 configurations. Does “OLS 1x1” mean that only the local-grid-cell CCFs are used as predictors, while the underlying data remain at the original $5^\circ \times 5^\circ$ resolution? Or, does it indicate that the $5^\circ \times 5^\circ$ results were regridded to a $1^\circ \times 1^\circ$ spatial resolution? Similarly, for the other configurations, does 7x5 describe the number of neighboring grid cells included in the regression—for example, seven grid cells in latitude, potentially covering 35° meridionally, and five grid cells in longitude, potentially covering 25° zonally—or does it describe the spatial resolution of the analysis? A clearer explanation of this nomenclature would be helpful.

Reply: Thank you for raising this. All grid cells in our analysis are $5^\circ \times 5^\circ$ (Sect. 2.1), and the “N×M” label gives the number of these $5^\circ \times 5^\circ$ grid cells spanned by the regression domain in latitude × longitude, not a spatial resolution or a regridding step. So RIDGE 3×3 means the target cell's low-cloud CRE anomaly is regressed on the CCFs from all 9 grid cells (3 cells in latitude by 3 in longitude, $15^\circ \times 15^\circ$ total) centered on and including the target cell, and OLS 1×1 is the special case using only the target cell's own local CCFs.

We agree this was only implicit in the manuscript, and have added a brief clarification to Sect. 3.1. The relevant sentence (L181–182) now reads:

“We test domains from 1×1 to 9×9 of these $5^\circ \times 5^\circ$ grid cells, using the CCFs from all cells in the domain as predictors while regressing against the low-cloud CRE anomaly of the center target cell only (e.g., a 3×3 domain uses CCFs from 9 grid cells, $15^\circ \times 15^\circ$, centered on the target cell).”

Perfect-model comparison

Below, I present an attempt to reproduce the perfect-model comparison as closely as possible following your approach. Unlike Park et al. (2025), in which the model CCF-predicted ERFaci was estimated by scaling the observationally derived CCF relationship, I performed a model-specific CCF analysis using each model's own aerosol, cloud, and meteorological variables. I used the 1995–2014 historical simulations to derive the susceptibility, consistent with the period used in your study. The CCF relationships were estimated using ordinary least squares (OLS) regression. Changes in aerosol concentration, represented by either aerosol index (AI) or sulfate mass concentration at 925 hPa (SO₄), were calculated as the difference between the preindustrial period (1850–1859) and the present-day period (2005–2014). All variables were spatially averaged to a $5^\circ \times 5^\circ$ resolution before performing the regression analysis. At present, I have analyzed only the first ensemble member (r1) from each model. Therefore, some of the models included in your study are not represented in the comparison below. The definition of ERFaci_{true} is similar to that of the ground-truth ERFaci used in your study. However, I

separated the cloud radiative contribution using radiative kernels and isolated the low-cloud contribution using the Webb et al. (2006) cloud classification.

Figure 1 illustrates how the choice of evaluation domain can affect the conclusions, particularly for AI. The upper panels show ERF_{aci} estimates averaged over land and ocean between 55°S and 55°N, whereas the lower panels show the corresponding results over ocean only. Because of the limited availability of ISCCP cloud fraction output, the analysis includes only six models for both AI and SO₄. For AI over the combined land+ocean domain, including the activation rate results in a slightly larger RMSE than excluding it (0.72 vs 0.68). Although the difference is smaller than that reported in your study, the qualitative conclusion is similar. However, when the analysis is restricted to ocean regions, the conclusion changes: accounting for activation produces a slightly lower RMSE than neglecting it (0.67 vs 0.68). For SO₄, accounting for the activation rate consistently produces a lower RMSE over both the combined land+ocean domain and the ocean-only domain. Moreover, restricting the analysis to ocean regions leads to a relatively large reduction in RMSE for the activation-rate case, from 0.58 to 0.50. In contrast, the RMSE for estimates without activation remains nearly unchanged when land is excluded, increasing slightly from 0.89 to 0.91. These results suggest that the inferred effect of accounting for aerosol activation may depend on both the aerosol proxy and the evaluation domain. The differences between our results may also arise from the definition of the ground-truth ERF_{aci} and the limited number of available models. With such a small sample, modest differences in the set of models included in the analysis could affect the resulting RMSE and the overall conclusion.

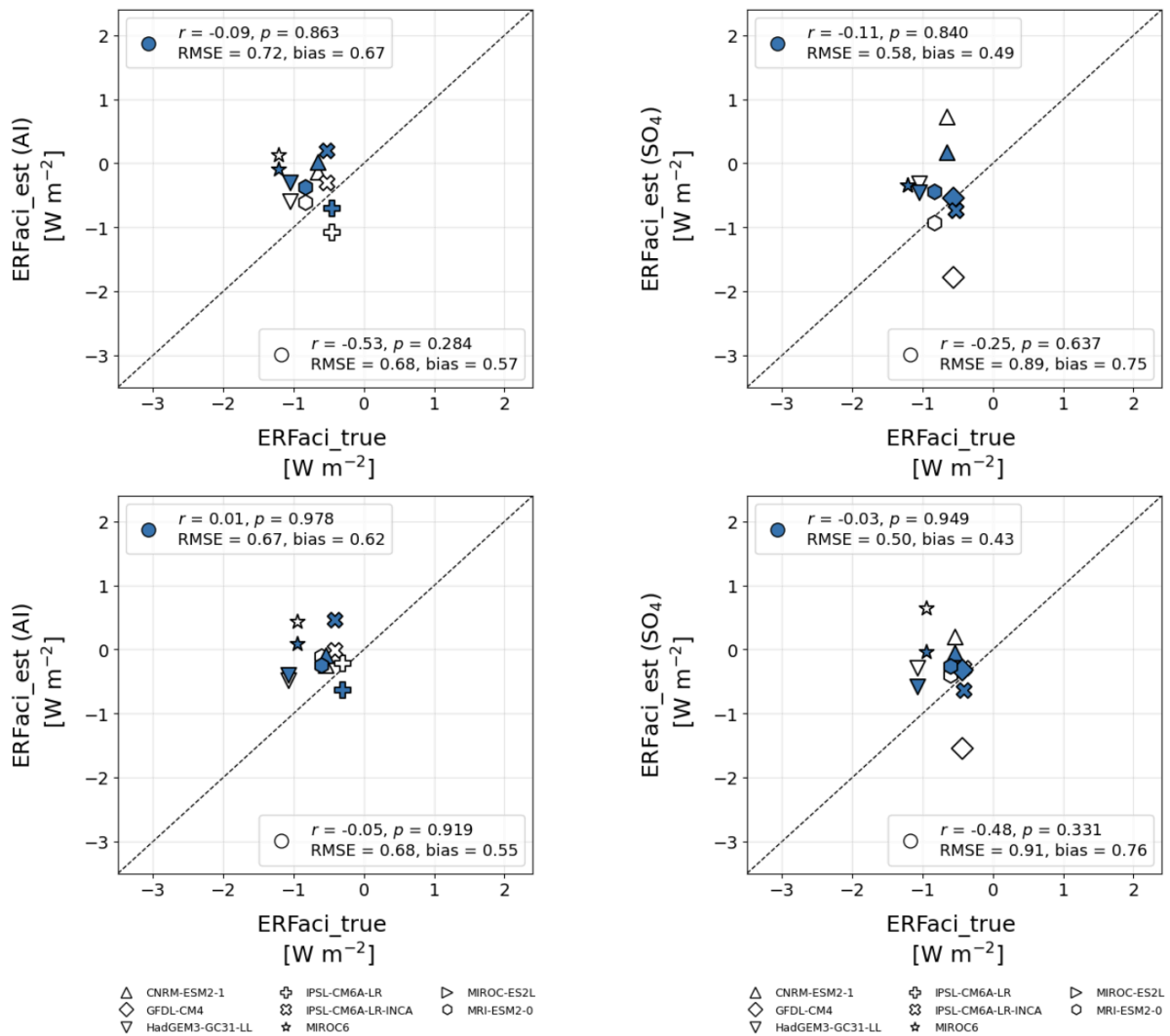


Figure 1. Perfect-model cross-validation of ERFaci estimates averaged over 55°S–55°N for land and ocean combined (upper panels) and ocean-only (lower panels). The left panels show the ground-truth ERFaci (ERFaci_true) versus ERFaci estimated using a cloud controlling factor analysis with aerosol index (AI) as the aerosol proxy. The right panels show the corresponding results using sulfate mass concentration at 925 hPa (SO₄). Filled blue markers represent estimates that account for the aerosol activation rate, whereas open hollow markers represent estimates that do not account for activation. The correlation coefficient (r), associated p-value, root mean square error (RMSE), and bias are shown in either the upper-left or lower-right corner of each panel. Bias is defined as the mean absolute deviation from the 1:1 line, shown by the dashed line.

To increase the number of models available for the perfect-model comparison, I examined whether ERFaci_est derived from cloud radiative kernel (CRK) estimates of non-obscured low-cloud radiative anomalies is consistent with ERFaci_est derived using the radiative kernel combined with the Webb et al. (2006) cloud classification. As shown in Fig. 2, the two

approaches produce generally consistent $ERFaci_est$ values for most models. The main exceptions are IPSL-CM6A-LR and IPSL-CM6A-LR-INCA, which show substantial deviations from the relationship found for the other models. Given the overall agreement, I used the radiative-kernel-derived low-cloud radiative anomalies to expand the sample to 10 models for Al and 12 models for SO_4 , while excluding the two IPSL models. This provides a larger model sample while maintaining consistency with the CRK-derived $ERFaci$ estimates.

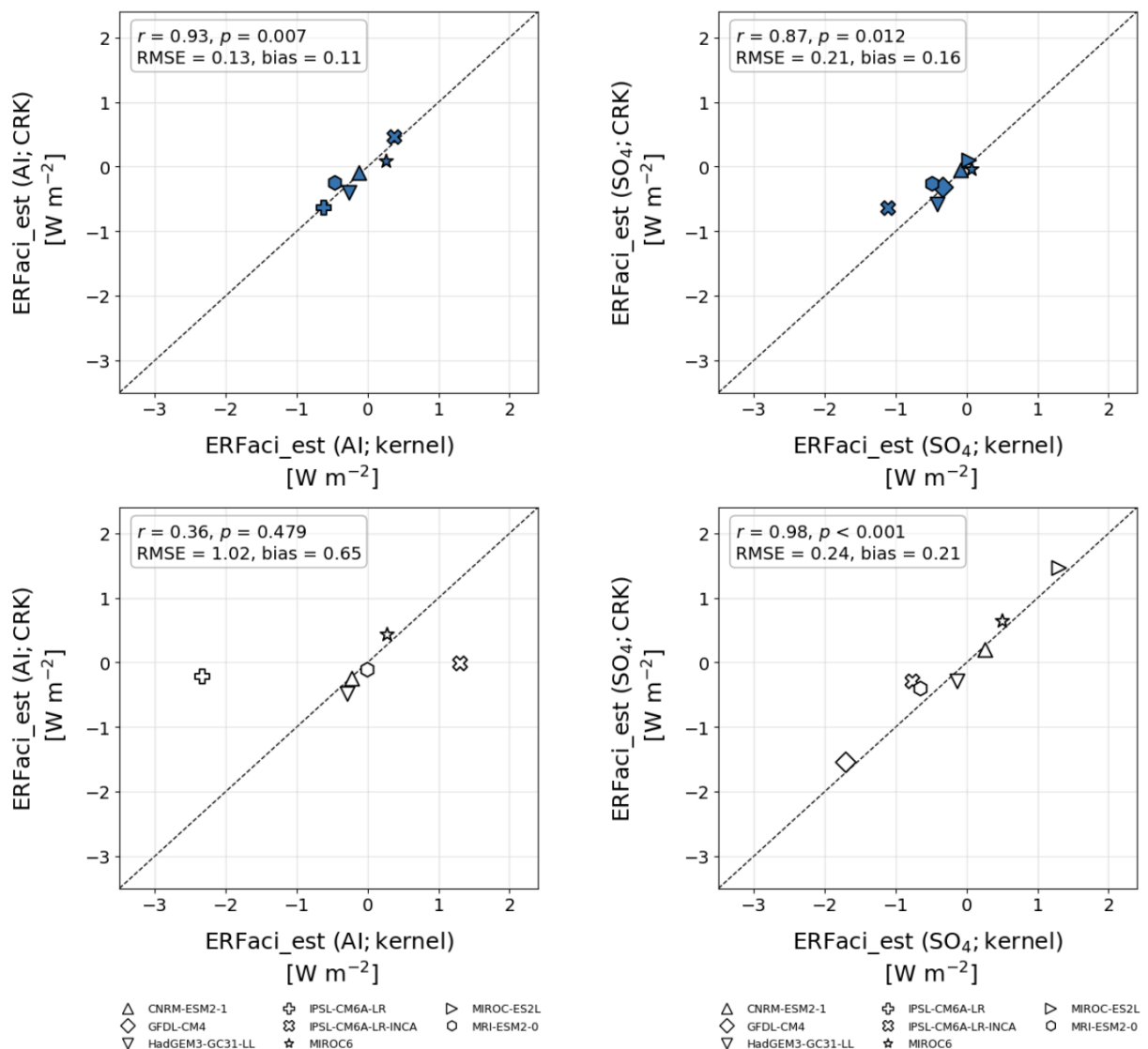


Figure 2. Comparison of $ERFaci_est$ calculated using two different definitions of low-cloud radiative anomalies. The y-axis shows estimates based on a cloud radiative kernel (CRK) applied to non-obscured low clouds, whereas the x-axis shows estimates based on the

radiative kernel combined with the Webb et al. (2006) cloud classification to isolate low-cloud radiative anomalies. All ERFaci_est values are averaged over the ocean-only domain from 55°S to 55°N. The upper row shows estimates that account for the aerosol activation rate, while the lower row shows estimates without activation. The left column uses aerosol index (AI) as the aerosol proxy, whereas the right column uses sulfate mass concentration at 925 hPa (SO4).

With the expanded model sample, we again find that accounting for the aerosol activation rate improves the perfect-model comparison. In Fig. 1, the AI-based analysis over the ocean-only domain showed little difference in RMSE between estimates that included activation and those that omitted it. However, after increasing the number of models in Fig. 3, the activation-included estimates clearly yield a lower RMSE. For SO4, the results are more robust across both the original and expanded model samples: estimates that account for activation consistently agree better with the ground-truth ERFaci than estimates that neglect activation. Overall, the expanded analysis further supports the conclusion of Park et al. (2025) that explicitly accounting for aerosol activation improves CCF-based ERFaci estimates in perfect-model comparisons. Although Park et al. (2025) used a simplified approach to estimate ERFaci_est, the present analysis further confirms that accounting for aerosol activation improves its relationship with ERFaci_true in aerosol-only forcing experiments. Because your study reaches the opposite conclusion based on a land-inclusive evaluation domain, a relatively small subset of models, and a ground-truth ERFaci estimate that does not isolate the low-cloud contribution, I believe the manuscript would benefit from a discussion of the possible reasons for this discrepancy. Thank you again for this valuable study. I would appreciate clarification of these points during the revision.

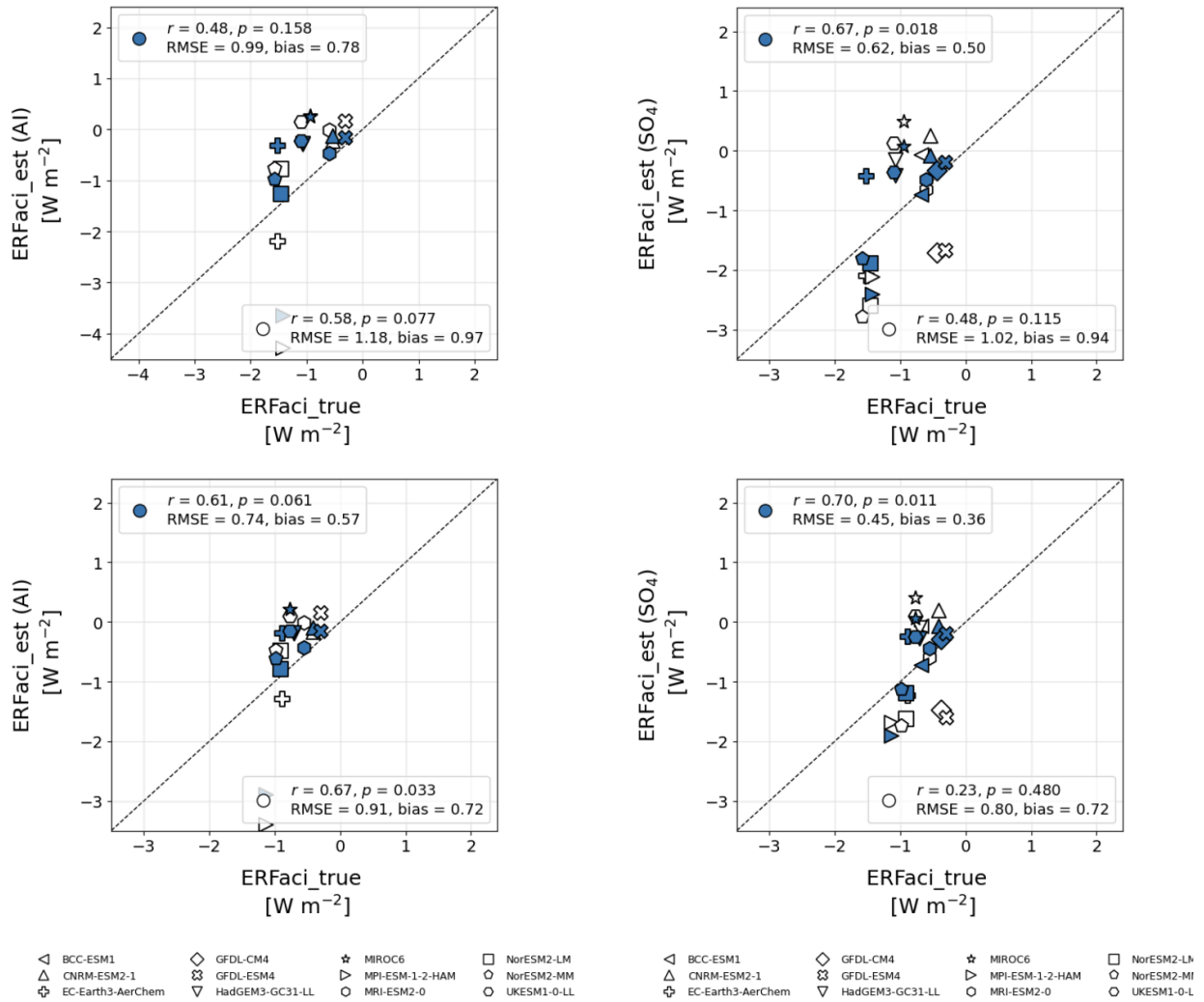


Figure 3. Same as Fig. 1, but using ERFaci_est derived from the radiative-kernel plus cloud-classification method in order to expand the number of models available for the perfect-model comparison. The upper panels show ERFaci estimates averaged over the ocean-only domain from 55°S–55°N. The lower panels show global ERFaci estimates obtained by scaling the 55°S–55°N ocean-only estimates by the ratio of global ERFaci_true to ocean-only 55°S–55°N ERFaci_true.

Reply: We want to thank the commenter again for this additional analysis. Reproducing our full evaluation pipeline independently, extending it to more models via a second ground-truth definition, and sharing the resulting figures in this much detail represents a substantial investment of time and effort.

As we described in our reply to Reviewer #2's 1st comment above, the smaller ensemble we use is not an incidental limitation but a direct consequence of a specific choice: our CMIP6 evaluation reproduces the same ISCCP-simulator-based, non-obscured low-cloud decomposition (Eq. 3 in the main text) used in our observational CERES pipeline, at every grid

cell, which restricts us to the 10 models with full ISCCP-simulator output. Our understanding is that the commenter's approach, and Park et al. (2025)'s own comparison, instead classify each grid cell into a single cloud regime following Webb et al. (2006) combined with a radiative-kernel decomposition, which is not the same quantity as the CERES pipeline it is meant to validate, but which does allow the sample to be extended further.

The commenter's own Fig. 2, comparing the CRK-based (ISCCP-simulator) $ERF_{aci,est}$ against the Webb-classification-based $ERF_{aci,est}$ for the standard (no-activation) route, is genuinely useful evidence on exactly this point. Looking at the standard-route panels there, the two ground-truth/pipeline definitions disagree by an RMSE of roughly 1.02 Wm^{-2} for AI and roughly 0.24 Wm^{-2} for SO4. These are of the same order of magnitude as the RMSE differences we (and the commenter) find between the standard and activation routes throughout this discussion, and as the total RMSE between the CCF analysis and the ground truth ERF_{aci} . Simply changing which pipeline is used to define " $ERF_{aci,est}$ " shifts the estimate by about as much as the activation-versus-standard question itself. We also note that the RMSE values in the commenter's Fig. 1 and Fig. 3 (using the Webb-classification ground truth) are considerably higher overall than the RMSE we obtain in our own low-cloud-only comparison using the ISCCP-simulator ground truth (referenced in our reply to Reviewer #2's third comment above), as well as our all-cloud comparison RMSE. We think this further supports our view that the specific method used to isolate the low-cloud signal is not a minor implementation detail, and that the ISCCP-simulator decomposition, being the one that mirrors our observational CERES pipeline term-for-term, gives the more faithful representation of what the observational analysis is actually measuring.

Taken together, we think this reinforces the concern we raised in our reply to Reviewer #2: once two evaluation pipelines diverge by this much on the standard route alone, a larger sample built by mixing pipelines is not necessarily more informative about the activation-versus-standard question than a smaller sample built on a single, consistent pipeline. This is because the ranking of methods can plausibly flip depending on which of the two similarly-sized pipeline discrepancies happens to dominate in a given subsample.

With that context, we'd like to summarize where our accumulated evidence stands. Beyond our original CMIP6 comparison, we have now run: (1) the same CMIP6 comparison restricted to a genuinely like-for-like, low-cloud-only ground truth (isolated via the same ISCCP-simulator decomposition on the RFMIP aerosol-only runs, described in our reply to Reviewer #2's third comment above); and (2) an extension of our opportunistic experiments to explicitly include the activation route (described in our reply to Reviewer #2's first major comment above), which uses real observed aerosol and meteorological trends and does not depend on CMIP6 model sampling at all. Across the original all-cloud perfect-model comparison, the low-cloud-only perfect-model comparison, and the opportunistic experiments, in both the ocean-only and combined ocean-and-land domains, we consistently find that the activation route does not improve predictive skill over the standard route.

Weighing this against the commenter's finding, we see two independent lines of evidence, each built on a single internally consistent pipeline, pointing the same direction, against one line of evidence built by combining two pipelines that disagree with each other by a comparable margin

to the effect in question. We believe our evidence is the stronger basis for our conclusion, though we want to be clear that we do not consider this fully settled: if a future CMIP7 phase provides more models with full ISCCP-simulator output, extending our specific comparison without needing to change pipelines partway through would be a valuable and fairly direct way to further test the activation-versus-standard question, and we would welcome that. For now, given the consistency of our own results across two different, independently constructed tests, we retain the standard route as our primary methodology.

We expanded the first caveat in our conclusions to include it:

“expanding the model ensemble in future phases of CMIP would further refine these constraints, and test the robustness of the methodological rankings identified here.”

References

Park, C., Soden, B. J., Kramer, R. J., L’Ecuyer, T. S., and He, H.: Observational constraints suggest a smaller effective radiative forcing from aerosol–cloud interactions, *Atmospheric Chemistry and Physics*, 25, 7299–7313, <https://doi.org/10.5194/acp-25-7299-2025>, 2025.

Scott, R. C., Myers, T. A., Norris, J. R., Zelinka, M. D., Klein, S. A., Sun, M., and Doelling, D. R.: Observed Sensitivity of Low-Cloud Radiative Effects to Meteorological Perturbations over the Global Oceans, *Journal of Climate*, 33, 7717–7734, <https://doi.org/10.1175/JCLI-D-19-1028.1>, 2020.

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., Boucher, O., Dufresne, J.-L., Nabat, P., Michou, M., Yukimoto, S., Cole, J., Paynter, D., Shiogama, H., O’Connor, F. M., Robertson, E., Wiltshire, A., Andrews, T., Hannay, C., Miller, R., Nazarenko, L., Kirkevåg, A., Olivié, D., Fiedler, S., Lewinschal, A., Mackallah, C., Dix, M., Pincus, R., and Forster, P. M.: Effective radiative forcing and adjustments in CMIP6 models, *Atmospheric Chemistry and Physics*, 20, 9591–9618, <https://doi.org/10.5194/acp-20-9591-2020>, 2020.

Wall, C. J., Norris, J. R., Possner, A., McCoy, D. T., McCoy, I. L., and Lutsko, N. J.: Assessing effective radiative forcing from aerosol–cloud interactions over the global ocean, *Proceedings of the National Academy of Sciences*, 119, e2210481 119, <https://doi.org/10.1073/pnas.2210481119>, 2022.

Webb, M. J., Senior, C. A., Sexton, D. M. H., Ingram, W. J., Williams, K. D., Ringer, M. A., McAvaney, B. J., Colman, R., Soden, B. J., Gudgel, R., Knutson, T., Emori, S., Ogura, T., Tsushima, Y., Andronova, N., Li, B., Musat, I., Bony, S., and Taylor, K. E.: On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles, *Clim Dyn*, 27, 17–38, <https://doi.org/10.1007/s00382-006-0111-2>, 2006.

Zelinka, M. D., Myers, T. A., Qin, Y., Chao, L.-W., Klein, S. A., Po-Chedley, S., Ma, P.-L., Wall,

C. J., Ceppi, P., and Gettelman, A.: Recent cloud trends and extremes reaffirm established bounds on cloud feedback and aerosol-cloud interactions, *Communications Earth & Environment*, <https://doi.org/10.1038/s43247-026-03461-8>, 2026.

References:

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., et al. (2020). Bounding global aerosol radiative forcing of climate change. *Reviews of Geophysics*, 58, e2019RG000660. <https://doi.org/10.1029/2019RG000660>

Carslaw KS, Lee LA, Reddington CL, et al. Large contribution of natural aerosols to uncertainty in indirect forcing. *Nature*. 2013;503(7474):67-71. doi:10.1038/nature12674

Ceppi, P., & Nowack, P. (2021). Observational evidence that cloud feedback amplifies global warming. *Proceedings of the National Academy of Sciences*, 118(30), e2026290118. <https://doi.org/10.1073/pnas.2026290118>

Ceppi, P., Myers, T. A., Nowack, P., Wall, C. J., & Zelinka, M. D. (2024). Implications of a pervasive climate model bias for low-cloud feedback. *Geophysical Research Letters*, 51, e2024GL110525. <https://doi.org/10.1029/2024GL110525>

Ceppi, P., Wilson Kemsley, S., Andersen, H., Andrews, T., Kramer, R. J., Nowack, P., Wall, C. J., and Zelinka, M. D.: Emerging low-cloud feedback and adjustment in global satellite observations, *Atmos. Chem. Phys.*, 26, 4153–4171, <https://doi.org/10.5194/acp-26-4153-2026>, 2026.

Dagan, G. (2026). Opposing transient and equilibrium effective radiative forcing from aerosol-cloud interactions. *Nature Communications*, 17. <https://doi.org/10.1038/s41467-026-72896-6>

Duran, B. M., N. J. Lutsko, and C. J. Wall, 2026: Aerosol–Ice–Cloud Interactions in a Perturbed Parameter Ensemble. *J. Climate*, 39, 4183–4203, <https://doi.org/10.1175/JCLI-D-25-0724.1>.

Grosvenor, D. P., Sourdeval, O., Zuidema, P., Ackerman, A., Alexandrov, M. D., Bennartz, R., Boers, R., Cairns, B., Chiu, J. C., Christensen, M., Deneke, H., Diamond, M., Feingold, G., Fridlind, A., Hünerbein, A., Knist, C., Kollias, P., Marshak, A., McCoy, D., Merk, D., Painemal, D., Rausch, J., Rosenfeld, D., Russchenberg, H., Seifert, P., Sinclair, K., Stier, P., van Diedenhoven, B., Wendisch, M., Werner, F., Wood, R., Zhang, Z., and Quaas, J.: Remote Sensing of Droplet Number Concentration in Warm Clouds: A Review of the Current State of Knowledge and Perspectives, *Reviews of Geophysics*, 56, 409–453, <https://doi.org/10.1029/2017RG000593>, 2018.

Park, C., Soden, B. J., Kramer, R. J., L'Ecuyer, T. S., and He, H.: Observational constraints suggest a smaller effective radiative forcing from aerosol–cloud interactions, *Atmospheric Chemistry and Physics*, 25, 7299–7313, <https://doi.org/10.5194/acp-25-7299-2025>, 2025.

Wall, C. J., Norris, J. R., Possner, A., McCoy, D. T., McCoy, I. L., and Lutsko, N. J.: Assessing effective radiative forcing from aerosol–cloud interactions over the global ocean, *Proceedings of the National Academy of Sciences*, 119, e2210481119, <https://doi.org/10.1073/pnas.2210481119>, 2022.

Webb, M. J., et al. (2006), On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles, *Clim. Dyn.*, 27, 17–38, doi:10.1007/s00382-006-0111-2.

Zelinka, M. D., T. Andrews, P. M. Forster, and K. E. Taylor (2014), Quantifying components of aerosol-cloud-radiation interactions in climate models, *J. Geophys. Res. Atmos.*, 119, 7599–7615, doi:10.1002/2014JD021710.