–Legend—
Referee text is in black
Author response text is in blue
— — — —

Referee #1

In this paper, a Sonabend-W et al. quantified the longwave radiative forcing of contrails at the top-of-atmosphere (TOA) based on GOES-16 satellite observations and ERA5 reanalysis, and estimated that the longwave radiative forcing of contrails is 46.9 GJ on average over the Americas. The authors apply causal inference to discern the effect of contrails while controlling for radiative and cloud confounders. The authors compared their results with CoCiP data, and illustrates how the longwave warming of contrails varies between day and night.

This method is plausible, but the results are unreasonable, so the authors should check the details to correct any potential mistakes. The paper might be accepted after addressing the following issues:

1. According to Fig. 10, the new method yields a longwave forcing of zero at noon, which is not reasonable. Unless there is no contrail at all at noon (which is untrue), the longwave forcing of contrails should not be zero. Furthermore, the surface temperature is higher at noon-time, so theoretically the longwave forcing of a contrail should be significant. Therefore, the CoCiP RF is more reasonable than that calculated by the new method.

Note: The authors added some discussions to address this issue, but it is hard to believe a zero longwave contrail forcing at noon. The unrealistic zero forcing at noon might be induced by issues in the regression process (see the next comment).

We thank the reviewer for their insistence on this point. In developing our response to this, we have performed an audit on our code and data, investigated further using the synthetic test data, and have contemplated and discussed amongst the co-authors some additional recommended practices for applying causal inference techniques to subgroup analysis (here, the subgroups are the GOES-ABI grid pixel data broken into their local hours to form diurnal progression).

In so doing we have:

A) *Uncovered a minor error in Table 1 coefficient values.*

Prior to manuscript submission, we fixed an off-by-one coding error which had resulted in approximately a 1% error on the magnitudes of $A_{H=11}$ values used as inputs to the causal regressions. At that time we regenerated all the manuscript figures, but unfortunately neglected to update the coefficients of Table 1 prior to submission. We have now updated the coefficients in this manuscript revision.

Please note, this was discovered in the course of our audit of the diurnal trends figure, but it did not in fact change the diurnal trend figure, because those causal regressions had already been regenerated with the correct $A_{H=11}$ values before the initial submission. Nor did it change the central estimate of 46.9 GJ/km longwave forcing over the Americas, because the fix only changed the central estimate by 0.01 GJ/km.

B) *Used the synthetic data to generate a "calibration curve" for the causal regressions, by systematically varying the magnitude of the ground truth effect size.*

Doing this has uncovered an attenuation effect in the causal inference regressions: the smaller the magnitude of synthetic longwave forcing, the more severely the oRF estimate will under-estimate the magnitude. When the synthetic ground truth is approximately 10GJ GJ/km there is a noise floor, for which the oRF regression estimates are erroneously centered near 0 GJ/km. We have added a new Appendix D to the manuscript with the calibration curve figure and discussion.

We believe the physical phenomena or modeling artifacts we previously described in the manuscript could still be playing a role in driving $oRF_{H=3}$ closer towards zero during midday hours than CoCiP's estimations. However, based on the confirmed magnitude of this effect in synthetic data, we now interpret this signal:noise ratio driven attenuation as the primary cause of the diurnal trends figure $oRF_{H=3}$ estimates being close to zero in the midday local hours.

Note that our aggregated Average Treatment Effect central estimate of 46.9 GJ/km longwave forcing is situated in a relatively robustly calibrated portion of the calibration curve; we have added discussion of the the possibility that this estimate could be revised upward in magnitude with future work developing improved calibration of such causal regressions. However, we have intentionally *not* revised the central estimate based on this calibration curve in the abstract or elsewhere in the manuscript, because we would like to perform a more thorough exploration of causal model calibration methodologies, deferred to future work.

C) *Determined that the diurnal trend figure is better placed in the appendix as an initial case study of causal inference subgroup analysis.*

Splitting out the GOES-16 ABI pixel gridded data into groups based on the local hour of their observation constitutes a form of subgroup analysis, and we apologize that we have not been

able to perform sufficient analysis to confidently separate potential modeling artifacts (such as uncontrolled confounding) apart from potential physical phenomena (such as atmospheric dehydration) in this context.

In consideration of the scope of the subgroup analysis methodology we now wish to be able to apply, thoroughly explore and validate (details in the new Appendix E), we would like to respectfully request the consent of the Reviewers and Editor to largely defer this type of diurnal subgroup analysis to future work.

1. In Eq. (3), a simple linear regression is used to calculate the parameters in the equation. However, as the authors pointed out, the correlation between independent variables Ai and Aj is large, so linear regression is not valid in this case. If the authors keep using simple linear regression, then this equation should be rewritten.

We appreciate the reviewer's comment and agree that if multiple linear regression with collinear input variables were being used it would not be valid, but our use of single linear regression is valid in this case because we're using a single advected trace density variable per regression fit. To make this clearer we have defined a new variable D_h which represents the trace density that has been advecting for h hours (rounding down partial hours), where h is in (0, …, 11). This lets us be more explicit about defining the cumulative average advected trace density as:

$$A_H = \frac{1}{H+1} \sum_{h \leq H} D_h, \text{ for } H = 0, \ldots, 11.$$

This notation allows differentiating between advected trace density that has advected for approximately h hours vs. the cumulative variables used in our regressions. If we were to use the hourly advected trace density D_h rather than A_H in Model (3), it would instead be the following:

$$E[\mathrm{OLR}_{\mathrm{COIN}} | \mathrm{CP}, D_0, \ldots, D_{11}, \mathrm{OLR}_{\mathrm{ERA5}}] = \alpha_0 \cdot \mathrm{OLR}_{\mathrm{ERA5}} + \sum_{j=0}^{4} I_{\{\mathrm{CP}=j\}} \left(\beta_j + \sum_{h=0}^{11} \gamma_{jh} D_h \right).$$

However, this model is not valid because it suffers from the high degree of correlation between different D_h as noted by the reviewer. For this reason we suspect the reviewer's comment is referring to correlations between D_i and D_j rather than between A_i and A_j and we apologize for the lack of clarity in our notation here. To be very explicit, in our method we only ever fit a single regression at a time with advected trace density treatment A_H from one maximum advection age H, where A_H is a scalar value per pixel. For example, we are never fitting a linear regression on multiple correlated inputs $A_{H=3}$ and $A_{H=4}$.

The approach we visualize in the violin plot figures with H on the x-axis (Fig 9 and 10 in the newly revised manuscript) is perhaps more similar to a sensitivity analysis or model comparison than to multiple linear regression; we are visualizing how the cumulative effect estimate evolves as the time window for advection expands. To this end, in the figures where these cumulative estimates are rendered, the Y-axis labels clearly describe them as a cumulative quantity, and we use violin plots to visually emphasize they are fitted with separate regressions. In this revision we have also augmented the Figure 10 legend to reiterate that it is a plot of a cumulative quantity and caution that the slope of the curve at any point may not be a good indicator of the marginal radiative forcing of contrails having that age.