Responses to Reviewers:

We thank the reviewer for their insightful comments and apologize for any confusion caused by our initial description in the manuscript and responses. We appreciate the opportunity to clarify and enhance our explanation and manuscript.

Major Comment

I do not understand the reply to my major comment 3 regarding whether g-computation and multiple linear regression results differ. From my read of section 3.4 describing the g-computation model and the Zenodo-archived code, the g-computation model produces the same output as the underlying multiple linear regression (the "Q-model"). If the multiple linear regression model is being used to estimate the aerosol effect on ETH, that is perfectly fine. However, if that is the case, the language in the study needs to be changed throughout it because it is stating that a new causal model is being used that is superior to past studies using regressions or other predictive models in isolating causal effects. From what I can tell, this isn't true, but it is possible I am missing something. If I am, then I suspect others will too because it is not clear from the paper or the response to my comment. Thus, I think it is important that the authors clearly demonstrate how the result from the g-computation model differs from that obtained from the multiple linear regression model alone with A=1 and A=0 inputs, which just produces the b1 coefficient being multiplied by A, as shown below.

As stated in section 3.4, a multiple linear regression (the Q-model) is performed with standardized variables ($Y = b0 + b1*A + b2*V1 + b3*V2$). Y is ETH, V1 and V2 are meteorological confounders, and A is aerosol concentration, where observations are inputted to derive the b coefficients. Then, 0 is used as an input for A to represent clean conditions and 1 is used as an input to represent clean conditions with V1 and V2 either held constant or entered as observed values for each observations (this isn't entirely clear, but shouldn't matter so long as V1 and V2 inputs are the same for both the clean (A=0) and polluted (A=1) calculations). The 2 multiple linear regression calculations are differenced to give a change in Y (ETH). Since A is standardized, A = 0 is the same as the mean aerosol concentration and A = 1 is 1 standard deviation above the mean and the change in Y (ETH) is associated with a 1 standard deviation change in A (aerosol concentration). Is this what is being done? If so, then this would just give b1 as the answer, which is the sensitivity of Y (ETH) to A (aerosol concentration). Mathematically, holding V1 and V2 to constant values or using the values from any given event to control for confounding variables, this would be:

$$Y(A=1) - Y(A=0) = (b0 + b1 + b2*V1 + b3*V2) - (b0 + b2*V1 + b3*V2) = b1$$

Weighting this result by polluted vs. clean conditions still gives b1 since it is a constant, and the ETH change becomes the change per 1 standard deviation change in aerosols (A=1 minus A=0) holding confounding factors the same for each scenario. Note that b1 is easier to interpret than the ETH change that is provided now because it is an ETH change per

aerosol concentration change. This is purely obtained from the multiple linear regression model and similar to past studies using multiple linear regression to estimate causal effects. Perhaps something else is being done, but this is how the paper and code currently read.

To clarify how calculations are being done, I suggest writing out the math so that it is clear how the ETH change is obtained to conclusively show one way or the other whether anything other than the multiple linear regression is being used. If only the regression is being used, then the paper should remove reference to g-computation and a new causal inference framework that is superior to past regression or random forest approaches. If instead something is being done that produces a different result than the regression alone, then that needs to be clearly demonstrated.

The reviewer is correct in recognizing that the coefficient *b1* from the standard multiple linear regression (MLR) model coincides with the Average Causal Effect (ACE) estimated through g-computation in the original manuscript. Like many models, g-computation is not without limitations, and using a simple, standard MLR model as the Q-model can obscure the inherent advantages of this causal framework, making the result quantitatively indistinguishable from *b1*, despite their *fundamental differences*.

In our response below, we demonstrate how employing more complex Q-models removes this coincidence. We therefore encourage readers to use more sophisticated Q-models when applying g-computation (if sample size allows).

In addition, the flexible nature of g-computation allows for the incorporation of machine learning models for resolving *non-linear* relationships within the data - something not achievable with standard MLR.

The reviewer's feedback also motivated us to explore a model-ensemble approach, where we report ACE distributions from g-computation using multiple Q-models for each scenario. This, too, is not feasible with MLR alone. We believe that this approach enhances the robustness of our findings, reduces biases associated with relying on a single model, and provides uncertainty estimates (which is rarely done in observational studies on aerosol-convection interactions).

We explain in detail below:

1. Results When Using Different Q-models

In the revised manuscript, we eliminated the use of the standard MLR and introduced three other Q-models to demonstrate the advantages of the g-computation and reveal the difference between *b1* and ACE. It also helps prevent confusion between standard MLR and g-computation.

1.1. First Q-Model: MLR with Interaction Terms

To *intuitively* illustrate how the result from g-computation differs *quantitatively* from the MLR coefficient *b1* (equivalent to β_1 in the equations below), we still use a MLR model but

with interaction terms to account for potential interactions between the exposure variable and confounders.

The outcome *Y* can be expressed as:

$$Y = \beta_0 + \beta_1 X + \beta_2 A + \beta_3 B + \beta_4 (A \cdot X) + \beta_5 (B \cdot X) \quad (1)$$

Here, ***X*** is the exposure variable, ***A*** and ***B*** are confounders, β_0 is the intercept, β_1, β_2, β_3, β_4, β_5 are coefficients, and ***A*·*X*** and ***B*·*X*** are the interaction terms.

Next, we generate counterfactual outcomes by setting the exposure variable ***X*** to fixed values representing different scenarios. As we demonstrate below, this step is critical for estimating the ACE and clearly sets this methodology apart from using MLR alone. We first forcefully set ***X*** = 1 for each individual case in the data for the polluted condition, and the potential outcome value that would have been observed is expressed as:

$$Y|do(X = 1) = \beta_0 + \beta_1 + \beta_2 A + \beta_3 B + \beta_4 A + \beta_5 B \quad (2)$$

We then forcefully set ***X*** = 0 for each individual case in the data to calculate the potential outcome that would have been observed if every case occurs in a clean condition:

$$Y|do(X = 0) = \beta_0 + \beta_2 A + \beta_3 B \quad (3)$$

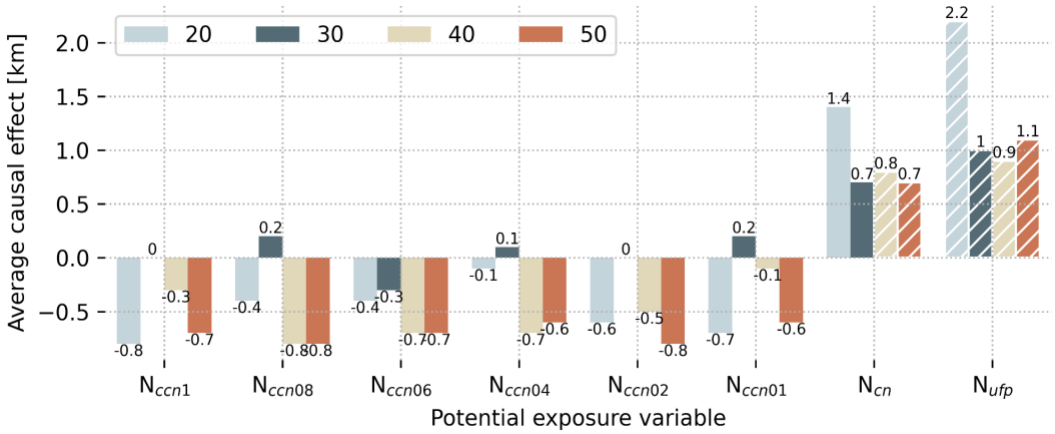
Therefore, the expected ACE of ***X*** on ***Y*** can be expressed as:

$$ACE = E\,[Y|do(X = 1) - Y|do(X = 0)] = \beta_1 + \beta_4 E[A] + \beta_5 E[B] \quad (4)$$

, where E[***A***] and E[***B***] are the expected values of ***A*** and ***B***, represented as the population means of confounders ***A*** and ***B***.

Due to the inclusion of interaction terms, the ACE (calculated using Equation 4 and shown in the figure below) *no longer* equals β_1 (*b1* originally), unlike in the simpler MLR example from our original manuscript. Although the differences between this ACE and the one presented in our original manuscript are within ±0.1 km (which may seem negligible for this study), g-computation uncovers causal effects rather than merely representing associations between variables, as MLR does.

We replaced the standard MLR with the Equation 1 in the revised manuscript and explicitly listed all equations above to help readers better understand g-computation.



1.2. Second Q-Model: Elastic Net Regression

The second Q-model we tested is the Elastic Net regression (Zhou and Hastie, 2005), an extension of linear regression that incorporates both Lasso and Ridge regularization penalties (two widely used regularization techniques in machine learning) into the loss function. This approach helps prevent overfitting and improves generalizability, which is especially important for small datasets like ours, where MLR models may struggle with overfitting. Note that the model equation for ***Y*** remains a MLR equation as shown in Equation 1 (with interaction terms), but the regularization penalties affect how the coefficients (β_0, β_1, β_2,…) are estimated.

The loss function can be expressed as:

$$L = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda[\alpha \sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2] \quad (5)$$

The first part, $\frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, represents the mean squared Error (MSE), measuring the prediction error which is the difference between predicted ($\hat{y}_i$) and actual values (y_i) (same as in MLR). The second part is the penalty term, which includes both the Lasso penalty (L_1) and the Ridge penalty (L_2). The Lasso penalty, $\sum_{j=1}^{p}|\beta_j|$, adds the sum of absolute values of coefficients as a penalty term and encourages sparsity by pushing some coefficients to zero. It helps reduce model complexity by eliminating irrelevant features/predictors. Meanwhile, the Ridge penalty, $\sum_{j=1}^{p}\beta_j^2$, adds the sum of squared coefficients as a penalty term and shrinks all coefficients, reducing their magnitude without setting them to zero. It encourages small, nonzero coefficients, reducing the impact of multicollinearity. Note that n represents the total number of training examples (data point) in the dataset, while p represents the total number of predictor variables (features) in the model.

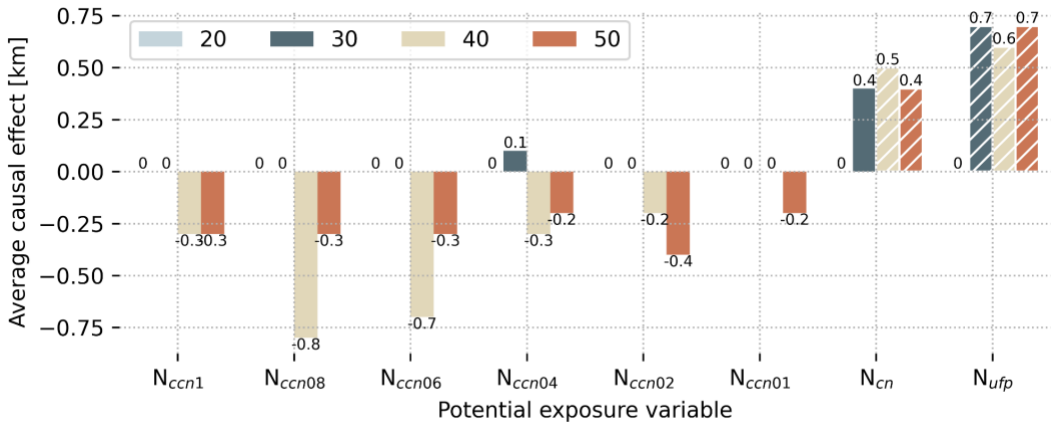
The parameter α is the mixing parameter (ranging from 0 to 1) that determines the balance between L_1 and L_2 regularization. For α = 0, the penalty is purely from Ridge regularization (L_2) and for α = 1, the penalty is purely from Lasso regularization (L_1). For 0 < α < 1, the

penalty is a combination of both L_1 and L_2, called Elastic Net regularization. There is no definitive way to choose this parameter, in our analysis, we set α to 0.5.

The parameter λ controls the strength of the regularization/penalty. A higher λ value increases the penalty, forcing coefficients β_j to be smaller, reducing overfitting. We fine-tune λ using a 5-fold cross-validation for each considered scenario separately (each bar in the figure below). Our results show that λ ranges from 0.01 to 10, depending on the scenario.

The ACE estimated using g-computation with Elastic Net regression as the Q-model is shown in the figure below. While the sign of the ACE for each scenario remains mostly the same as in the figure above, the overall ACE values are generally smaller, with some dropping to zero in certain scenarios. These zero ACE scenarios indicate that aerosols have no effect on ETH under these conditions. Note that the ACEs presented in the figure below also no longer equal β_1 (b1 in the original manuscript).

Section 4.3 of the revised manuscript presents the results from the Elastic Net regression Q-model, with details on the model setup and parameters provided in the supplemental material.



1.3. Third Q-Model: Support Vector Regression (SVR)

We further explore a non-linear Q-model using SVR (Smola & Schölkopf, 2004). SVR is a supervised machine learning technique designed to find a function such that most data points lie within an ϵ-tube, meaning their predicted values deviate at most ϵ from the true values. In other words, SVR aims to fit the data within a specified margin of tolerance (ϵ), balancing smoothness with accuracy. This approach penalizes only large errors that exceed ϵ, while small deviations are allowed.

For a training set $T = \{(X_i, y_i)\}_{i=1}^{l}$, where $X_i \in R^N, y_i \in R$, the SVR function can be expressed as:

$$f(X) = w^T \cdot \phi(X) + b \quad (6)$$

where **w** $\in R^N$ is the coefficient vector in feature space, $b \in R$ is the intercept, and $\phi(.)$ denotes the kernel function that maps the input ***X*** to a vector in the feature space. SVR supports multiple kernel functions (e.g., linear, polynomial, or Radial Basis Function) to model diverse data patterns. In our study, we use the Radial Basis Function kernel to capture potential non-linear relationships in our data.

The solution of **w** and *b* can be obtained by solving the optimization problems:

$$Minimize \ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) \ (7)$$

This is subject to the following constraints:

$$y_i - f(X_i) \leq \epsilon + \xi_i^+ \quad (8)$$

$$f(X_i) - y_i \leq \epsilon + \xi_i^- \ (9)$$

$$\xi_i^+, \xi_i^- \geq 0 \quad (10)$$

Here, the first term, $\frac{1}{2}||w||^2$, controls model complexity (smaller weights lead to a simpler model). The second term, $C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-)$, measures the sum of slack variables (ξ_i^+, ξ_i^-), which account for deviations beyond the margin ϵ. This term is introduced because a perfect fitted function $f(X)$ in Equation 6 with ϵ precision may not exist or feasible (Cortes and Vapnik, 1995). In other words, these slack variables allow the model to tolerate some degree of error in the fitting process.

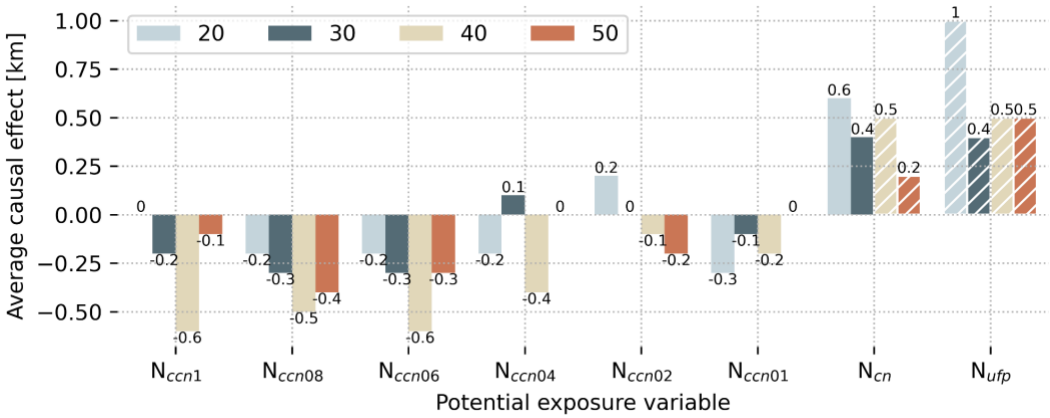
The parameter C (> 0) in Equation 7 regulates the trade-off between $f(X)$ complexity and the degree to which deviations from ϵ are tolerated. If C is too high, the model penalizes errors heavily and may overfit, trying to fit all points exactly; if too low, the model ignores smaller deviations and focuses on capturing general trends, leading to a simpler model. In our study, we set C = 1.0, which represents moderate regularization, balancing error minimization with model complexity.

The parameter ϵ defines a margin of tolerance within which errors are ignored; the model does not penalize errors smaller than ϵ. If ϵ is too small, the model captures more details but risks overfitting; if too large, it may miss significant variations in the data. We use 5-fold cross-validation to evaluate different ϵ values for each scenario and select the one that minimizes the mean squared error. Depending on the scenario, ϵ ranges from 0.1 to 1 in our study.

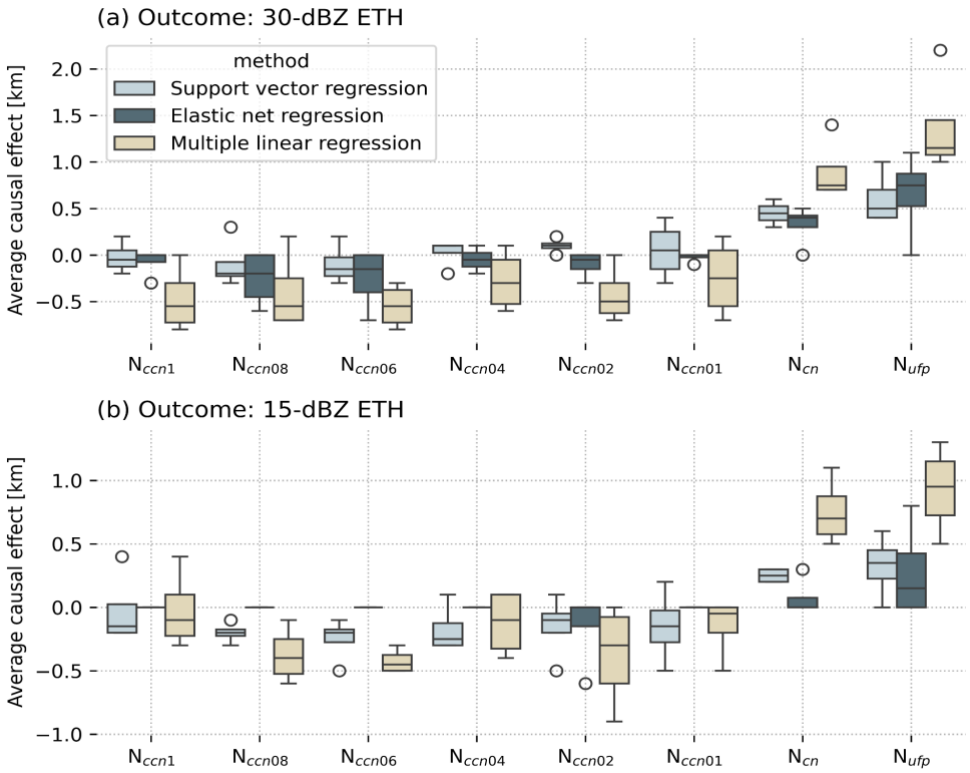
The results from the SVR-based Q-model are presented in the figure below. Overall, the estimated ACEs are smaller compared to those in our original manuscript. For a portion of

the scenarios, the values are reduced by half compared to the original results. However, the overall sign of the ACE remains consistent with the results from other Q-models.

Section 4.3 of the revised manuscript presents the results from the SVR-based Q-model, with details on the model setup and parameters provided in the supplemental material.



We present box-whisker plots to show the distributions of ACEs for all scenarios based on the three Q-models when using different exposure variables, as shown in the figure below. The differences are evident across the Q-models when both 30 dBZ and 15 dBZ are used as outcomes. Using MLR models (with interaction terms) overestimates the impacts (both positive and negative) of aerosols compared to the other two models. This result further reveals the limitation of aerosol impacts on DCC intensity.



Caption: Average causal effects on ETH estimated using three Q-models. The post-sounding aerosol averaging period is considered. Scenarios with different radii from the ARM sites are all included. The meteorological variables are calculated using ARM soundings (6-hr) when assuming the most-unstable parcel would rise to form a convection.

It is important to note that our candidate Q-models are somewhat constrained by the small sample size in this study. With a larger dataset, more complex Q-models, including a variety of machine learning-based predictive models, could be employed. The same subsequent steps can then be followed to calculate ACE. An ensemble modeling approach may be particularly useful in such cases (as shown in the figure above) to provide uncertainty estimates and enhance the robustness of the results. This is also critical for research on aerosol-deep convection interactions, where relying on a single model may introduce biases that skew conclusions in one direction or another (invigoration, enervation, or no effect). Employing multiple Q-models provides a way to quantify uncertainties and encourages researchers to use non-linear models and move beyond simple MLR.

Although we have demonstrated that the results from g-computation can be quantitatively different from those of MLR, we want to strongly emphasize that the two methods are *fundamentally different* in their ability to resolve true causal effects. MLR does NOT inherently account for confounding or causal structure, which is a key capability of g-computation. This distinction is particularly important in aerosol-deep convection interaction studies, where meteorological factors can often buffer aerosol effects (Stevens and Feingold, 2009). Properly accounting for confounders is essential for obtaining more accurate estimations of aerosol impacts on deep convection.

2. Summary of the differences between G-Computation and MLR

We recognize that when using a standard MLR, as in our original manuscript, the regression coefficient of the exposure variable quantitatively aligns with the results from g-computation. However, this does not mean that MLR and g-computation are equivalent; in fact, their purposes and interpretations are fundamentally different. Moreover, it also does not imply that MLR can be used for estimating causal effects without specific constraints or conditions. Additionally, g-computation is just one of many causal inference methods. If other causal inference methods/frameworks (e.g., propensity score matching [Rosenbaum et al., 1983]) were employed (not relying on fitting a Q-model), the results would probably differ between the two approaches (Chatton et al., 2020).

Basically, MLR is a statistical tool for modeling the relationship between a dependent variable and multiple predictors. Its primary purpose is to predict outcomes and estimate *associations*, not causal effects. It does *not* inherently account for confounding, and its regression coefficients only represent causal effects under *strict assumptions* (e.g., no unmeasured confounding, correct model specification, random exposure assignment).

G-computation, on the other hand, is a powerful causal inference method that explicitly estimates the causal effect of an intervention by modeling the relationship between variables and simulating *counterfactual* outcomes. G-computation controls for confounding by keeping confounding variables constant across hypothetical interventions (i.e., setting all exposure to 1 or 0). This ensures that the estimated causal effects are not biased by confounders, leading to a more accurate assessment of the true causal effect of the exposure on the outcome. Its flexible framework allows for the incorporation of advanced *non-linear* predictive models as Q-models to improve reliable and robustness of the estimation.

Finally, we encourage the use of more complex Q-models and a model-ensemble approach to fully leverage the advantages of g-computation.

3. Changes to the revised manuscript

We replaced the simple, standard MLR model with the first Q-model we presented in this document, and we added the Equations (1) - (4) to the revised manuscript (in section 3.4) to better explain the calculation. We uploaded the code to Zenodo. We also updated figures 8 - 11 and Tables 3 and S4 using results from new Q-models and model ensembles. We added the box-whisker plot shown above to the revised main manuscript as Figure 12. We also explicitly explain each additional Q-model and their parameter setting in the supplemental material. We made minor edits throughout the manuscript to reflect these changes.

Particularly, we added these paragraphs to the section 4.3 in the revised manuscript and add details of each Q-model to the revised supplemental materials:

"*The flexible nature of g-computation allows for the incorporation of advanced predictive models, such as machine learning models, to capture non-linear relationships within the data. Therefore, we employ the Support Vector Regression (SVR) model (Smola & Schölkopf, 2004) as the Q-model to examine whether the results change. Additionally, we explore a model-ensemble approach by reporting statistical results from g-computation using multiple Q-models for each scenario. To achieve this, we add results from the Elastic Net Regression Q-model to provide uncertainty estimates. We believe this approach enhances the robustness of our findings and reduces uncertainty associated with relying on a single model. This is also critical for research on aerosol-DCC interactions, where relying on a single model may introduce biases that skew conclusions in one direction or another (invigoration, enervation, or no effect).*

SVR is a supervised machine learning technique designed to find a function such that most data points lie within an ϵ-tube, meaning their predicted values deviate at most ϵ from the true values (Vapnik, 2013). In other words, SVR aims to fit the data within a specified margin of tolerance (ϵ), balancing smoothness with accuracy. This approach penalizes only large

errors that exceed ϵ, while small deviations are allowed. It is suitable for our small sample size. The detailed model and parameter settings are described in the supplemental material.

Elastic Net regression (Zhou and Hastie, 2005) is an extension of linear regression that incorporates both Lasso and Ridge regularization penalties (two widely used regularization techniques in machine learning) into the loss function. This approach helps prevent overfitting and improves generalizability, which is especially important for small datasets like ours, where MLR models may struggle with overfitting. Note that the model equation for Y remains a MLR equation as shown in Equation 1 (with interaction terms), but the regularization penalties affect how the coefficients (β_0, β_1, β_2,…) are estimated. We present the detailed model and parameter settings in the supplemental material.

We show box-whisker plots to illustrate the distributions of aerosol causal effects based on all three Q-models when using different exposure variables in Figure 12. The differences are evident across the Q-models when both 30 dBZ and 15 dBZ are used as outcomes. Using MLR models (with interaction terms) overestimates the impacts (both positive and negative) of aerosols compared to the other two models. This finding underscores the uncertainties in aerosol effects on DCC intensity and highlights the advantages of an ensemble modeling approach for providing uncertainty estimates and enhancing result robustness. This is particularly critical for aerosol–deep convection interaction research, where reliance on a single model may introduce biases that skew conclusions toward invigoration, enervation, or no effect.

Note that when using a standard MLR model (without interaction terms) as the Q-model for g-computation, the regression coefficient of the exposure variable from the standard MLR aligns quantitatively with g-computation results, potentially obscuring the latter's inherent advantages. However, this coincidence does not imply equivalence between MLR and g-computation; their purposes and interpretations are fundamentally different. Moreover, it does not suggest that MLR can estimate causal effects without specific constraints (e.g., no unmeasured confounding, correct model specification, random exposure assignment). To fully leverage g-computation's advantages, we encourage the use of more complex Q-models and a model-ensemble approach."

Minor Comment

Related to my previous major comment 1, I still feel that the introduction focuses primarily on aerosol effects on cloud dynamics. This is fine if the authors prefer to keep it this way, but why I had recommended broadening the discussion to direct aerosol effects on cloud microphysics is that throughout the introduction, there are references to aerosol-DCC interactions in general. For example, lines 31 and 41 begin paragraphs by using this phrase but then just focus on indirect effects on dynamics. Aerosol-DCC interactions is not synonymous with aerosol indirect effects on cloud dynamics, so I think it needs to be clearer

when mentioning aerosol-DCC interactions that the manuscript specifically focuses on aerosol effects on dynamics rather than direct effects of aerosols on hydrometeor properties or convective cloud effects on aerosols.

We agreed and we modified the first sentence of lines 31 and 41 to highlight that we are particularly interested in convective dynamics other than other aspects of the aerosol–cloud interactions in DCCs.

"*Aerosol–cloud interactions in DCCs, particularly the aerosol effects on convective dynamics, are among the most complex and challenging processes to simulate accurately. Note that this study and its introductory discussion mainly focus on aerosol effects on convective dynamics.*"