Responses to Reviewer #2

The authors present a machine learning based approach for disentangling the effects of different anthropogenic forcing agents on historical temperatures by training on an opportunistic ensemble of climate model simulations. They find a significant cooling contribution due to aerosol, which leads to enhanced future warming as aerosol emissions decline. The study is framed well and, if true, would represent an interesting and useful approach for attributing observed warming trends without having to run dedicated single (or all-but-one) forcing simulations.

That said, I have strong concerns about the methodology, and in particular the approach for training and testing the model. Because the scenarios used for training the model are all strong correlated, the use of a random sub-sampling of test data leads to a serious risk of overfitting. That is, the randomly sampled test data provides no real validation that the model is able to extrapolate to the scenarios the authors then use the model to explore, undermining the presented results. A more appropriate approach to select test data would be to hold back a whole scenario, as in ClimateBench (Watson-Parris et al. 2022), which also specifically tests emulators against ssp245-aero to perform aerosol attribution.

If invited for resubmission, the manuscript would benefit from proof-reading by a native English speaker as there are many grammatical and style aspects that could be improved.

Response：

We sincerely thank the reviewer for their insightful and constructive comments. We fully understand the reviewer's concern regarding the training and validation strategy, particularly the potential risk of overfitting due to strong correlations among different forcing scenarios. As the reviewer rightly pointed out, using randomly sampled data for validation may lead to overly optimistic performance estimates, as the test data may share similar temporal or forcing patterns with the training data, thus weakening the ability to assess the model's true generalization performance.

To address this critical issue, we have made a fundamental revision to our

training-validation approach. Specifically, we no longer use random sub-sampling to construct the test dataset. Instead, we now exclude the entire SSP2-4.5 scenario from the training data and use it solely as an independent validation set. This stricter partitioning can better evaluate the model's ability in generalizing the "unseen" scenarios and aligns with best practices recommended in the ClimateBench framework (Watson-Parris et al., 2022). This modification significantly reduces validation bias caused by inter-scenario correlation.

The LightGBM continues to perform well for the model testing of the held-out SSP2-4.5 scenario. The predicted global annual mean surface air temperature (GSAT) closely matches the CMIP6 multi-model ensemble mean, with an R^2 of 0.94, RMSE of 0.23 ℃, and MAE of 0.18 ℃ (Figure R1a). The model also demonstrates strong performance in reproducing regional and zonal surface air temperature, with R^2 values exceeding 0.90 across latitude bands in the Northern Hemisphere, and R^2 values between 0.7 – 0.9 in the mid- to high-latitudes of the Southern Hemisphere, as well as over China, Europe, and North America (Figures R1b–d and R2). Moreover, the model successfully captures the temporal evolution of global and regional mean temperatures from 2021 – 2100 under the SSP2-4.5 scenario (Figures R3–R6).

Moreover, using the newly trained model validated through the SSP2-4.5 scenario, we re-conducted the temperature attribution analysis. The results remain consistent with our original findings, which demonstrates good generalization and robustness under a more rigorous validation framework. We have revised the methods and corresponding results in the manuscript:

"The following steps describe the specific procedure by which ML models attribute and predict SAT (Fig. 1):

First, the datasets for training the machine learning models are constructed following the experimental design of CMIP6 simulations, with forcing factors varying according to the specific experiments and time period. For example, the historical experiment corresponds to the period of 1850–2014, during which all forcing factors vary with time. The hist-aer experiment simulates changes in anthropogenic aerosols from 1850 to 2020, where anthropogenic emission data of aerosols and precursors from

1850 to 2014 are derived from the historical inputs and data from 2015 to 2020 are based on the SSP2-4.5 scenario, while other forcing factors are fixed at the 1850 levels. Similarly, the hist-GHG, hist-nat, and hist-CO_2 experiments simulate individual changes in greenhouse gases, natural forcings, and CO_2, respectively, with other forcings held constant at 1850 levels. Future attribution simulations including ssp245-aer and ssp245-GHG represent variations in anthropogenic aerosols and GHGs, respectively, from 2015 to 2100, with other forcings fixed at 1850 levels. The SSP1-1.9, SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5 scenarios simulate concurrent variations in all forcing factors during 2015–2100. This dataset construction enables the machine learning models to leverage time-evolving forcing factor data from multiple experiments, facilitating accurate prediction and attribution of surface air temperature responses.

Secondly, one ML model is trained to predict GSAT, eighteen models are developed for zonal SAT bands from 90° S to 90° N, and three regional models focus on key regions including China, Europe and North America. **The training dataset combine data from historical, hist-GHG, hist-aer, hist-CO_2, ssp245-GHG, ssp245-aer, SSP1-1.9, SSP1-2.6, SSP3-7.0, and SSP5-8.5 experiments, while the SSP2-4.5 scenario is reserved as an independent test set to evaluate model performance, similar to the method recommended in the ClimateBench framework (Watson-Parris et al., 2022).** Key hyperparameters including boosting_type, objective function, num_leaves, num_boost_round, learning_rate, reg_alpha, reg_lambda, and colsample_bytree are optimized through five-fold cross-validation for each LightGBM model. The best performing hyperparameters for the GSAT model are: boosting_type = gbdt, objective = regression, num_leaves = 31, num_boost_round = 200, learning_rate = 0.05, reg_alpha = 0.1, reg_lambda = 0.1, and colsample_bytree = 0.9. The coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) are calculated to evaluate the performance of the ML models."

We believe that this approach, which trains machine learning models on multi-model simulation data while validating against an independent scenario, provides a more reasonable methodological pathway for rapid and quantitative attribution of

historical and future climate change without the need to run large numbers of single-forcing simulations. **Finally, in response to the reviewer's suggestion regarding language issues, we have carefully proofread and revised the manuscript to improve the accuracy and clarity of the English writing.**

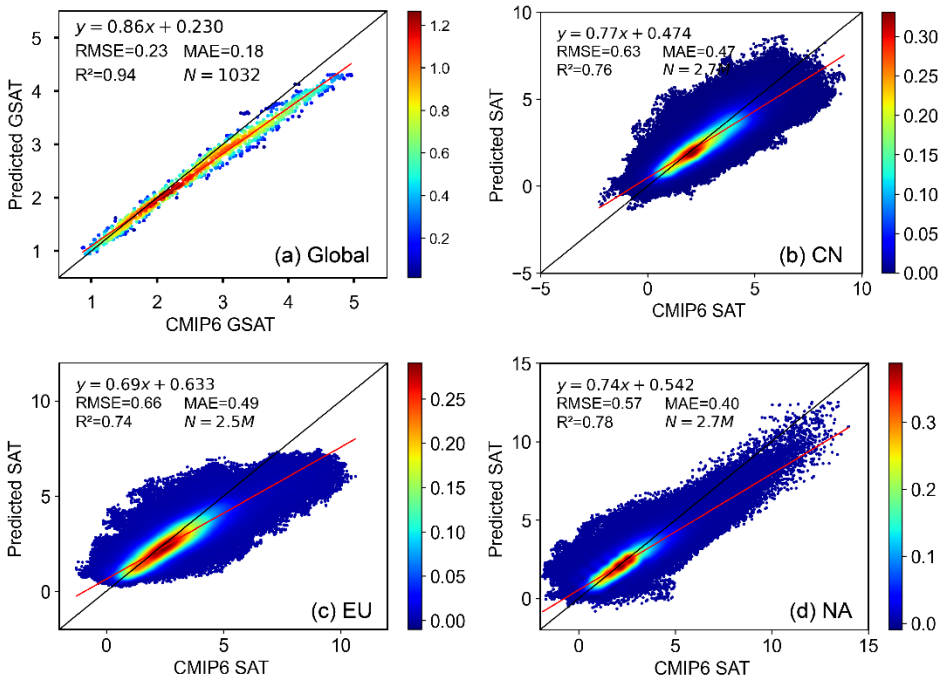

Figure R1. Scatterplot of the density of global and regional SAT (°C) over China, Europe, and North America from CMIP6 multimodel mean versus the predicted values from the LightGBM model under the SSP2-4.5 scenario. The black and red solid lines are the 1:1 lines and linear regression lines, respectively. Statistical metrics including RMSE, MAE, and R^2 are given in the upper left corner of each panel.

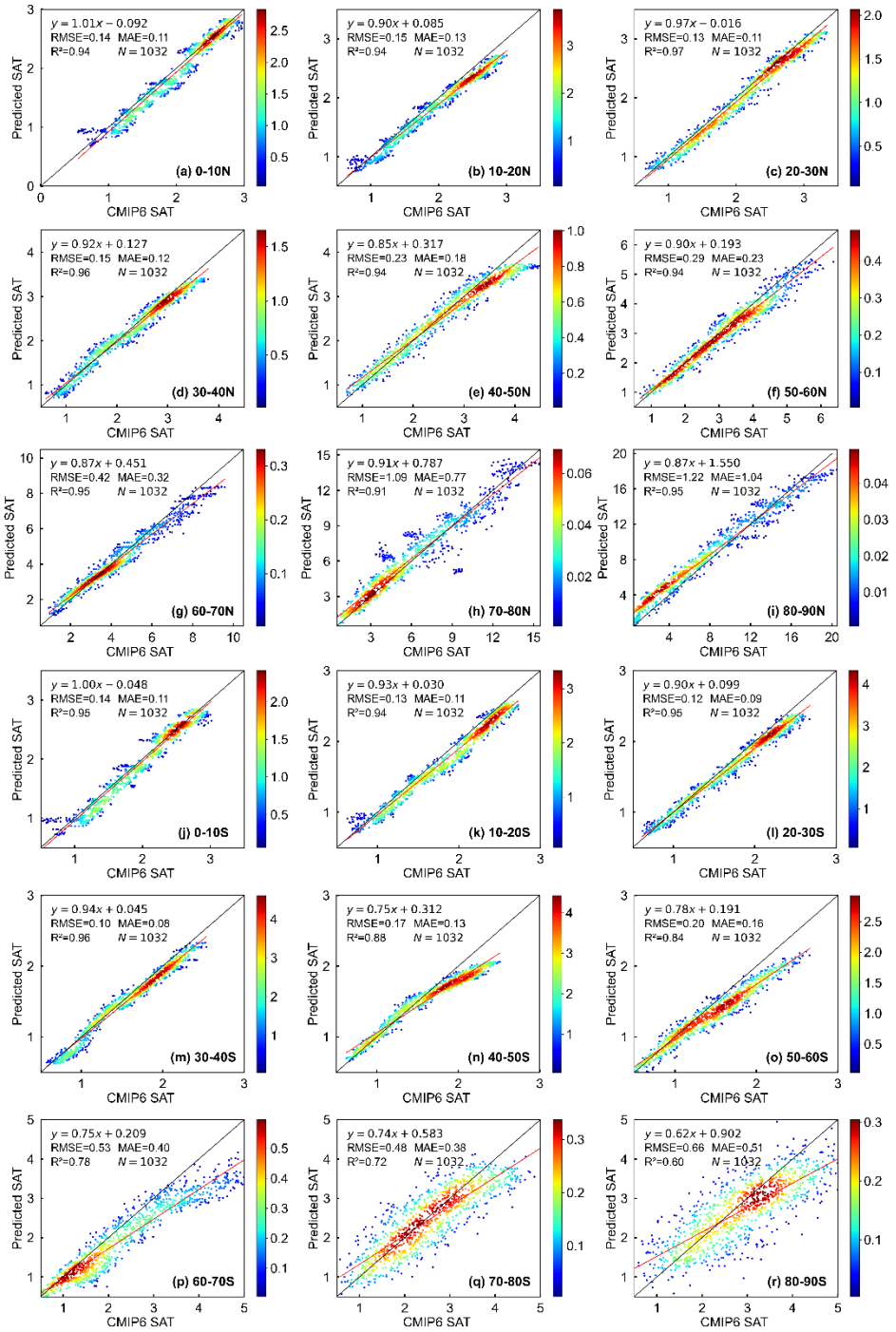

Figure R2. Scatterplot of the SAT density (°C) from CMIP6 multimodel simulations versus the predicted values from the LightGBM model for the eighteen latitudinal bands each spacing 10° from 90° S to 90° N, with color bars indicating the density of the data distribution. The black and red solid lines are the 1:1 lines and linear regression lines, respectively. Statistical metrics including RMSE, MAE, and R^2 are given in the upper left corner of each panel.

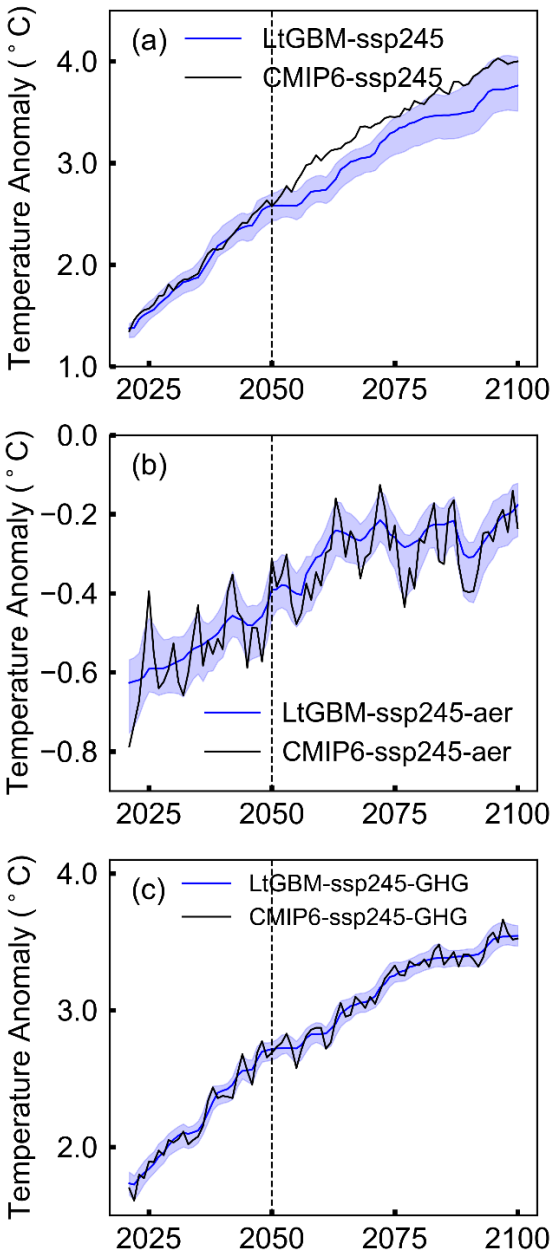

Figure R3. Time series of LightGBM-predicted GSAT anomalies (°C) and corresponding CMIP6 DAMIP values during 2021 – 2100 under the SSP2-4.5 scenario due to changes in (a) all forcing, (b) anthropogenic aerosols and (c) GHGs. Shaded areas indicate the range of the ML prediction by random separation of training and testing datasets for 100 times.

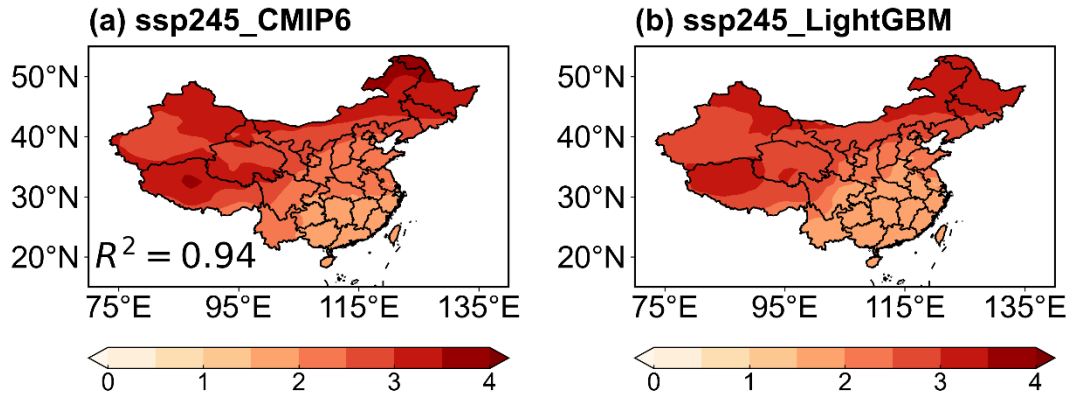

Figure R4. Spatial distribution of the 2015–2100 mean surface air temperature under the SSP2-4.5 scenario predicted by LightGBM and simulated by the CM IP6 multi-model ensemble.

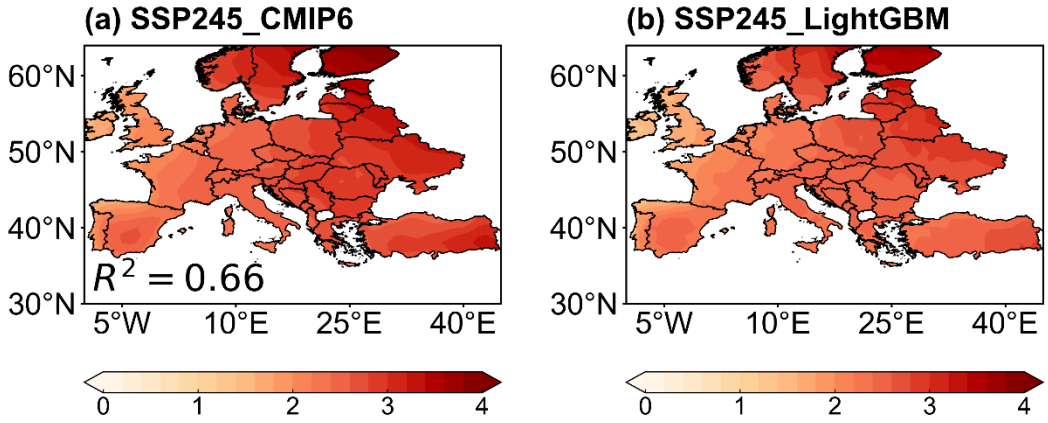

Figure R5. Same as Fig.R4, but for Europe.

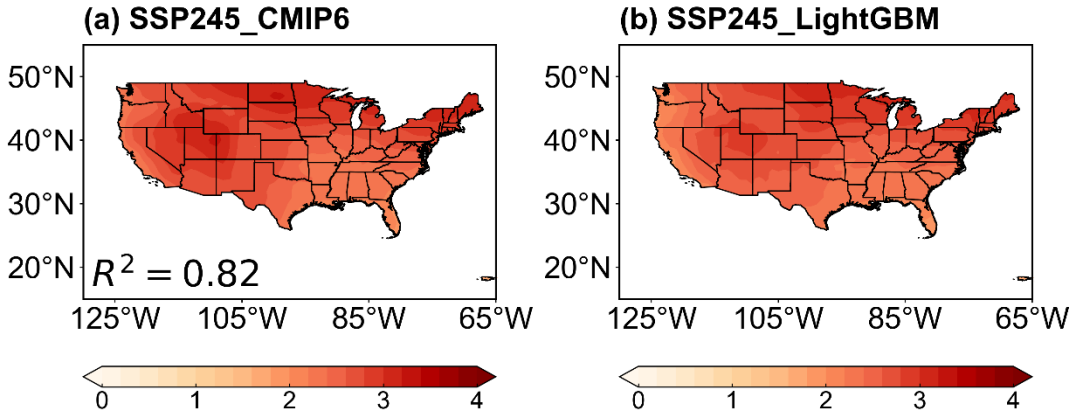

Figure R6. Same as Fig.R4, but for North America.

Reference:

Watson-Parris, D., Rao, Y., Olivie, D., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., Gonzalez, J., Harder, P., Jeggle, K., Lenhardt, J., Manshausen, P., Novitasari, M., Ricard, L., and Roesch, C.: ClimateBench v1.0: A benchmark for data-driven climate projections. Journal of Advances in Modeling Earth Systems, 14, e2021MS002954. https://doi.org/10.1029/2021MS002954, 2022.