Responses to Reviewer #1

This work deploys a decision tree based ML to generate an emulator for CMIP6-like climate model to attribute the contribution of anthropogenic aerosol and GHGs in global and regional warming. The ML-attribution is done for SSP119 and SSP585 scenarios. Although the attribution has been explicitly simulated in CMIP6-DAMIP for SSP245, the SSP119 (Net-Zero) and SSP585 (high fossil fuel) scenarios have no multiple model detection & attribution modelling experiments, therefore is of particular interest to the community. While interesting, I feel there are major concerns that this study should be addressed accordingly, before can be accept for publishing. I will provide details of my concerns as below, in addition, I feel the language presentation would also need further polishing.

We thank the reviewer for the helpful comments. Below, please see our point-by-point response (in blue) to the specific comments and suggestions and the changes that have been made to the manuscript, in an effort to take into account all the comments raised here.

Specific comments:

1. I doubt about the "perfect" validation of the ML model, eg. Fig.2a. I think the approach for model validation is not a fair way to do it. If I understand the method section correctly, the authors randomly leave-out 10% data from a bunch of CMIP6 models outputs for validation and use the rest 90% data for training. Therefore, the ML, in the majority cases, is only doing an easy interpolation job, rather than the challenging predicting job which we expected ML to do. Because, for example, CESM: 2010.01 – 2010.07 + 2010.09 – 2010.12 are used for training, but 2010.08 is used for validation (just an example, similar cases can happen in most of situations). In addition, your training dataset could still have the 2010.08 data from ECHAM, MRI-ESM, MPI-EMS, etc. Therefore, your ML model only learns the relationship between different models and interpolation between different months (or from neighbor grid cells). However, for

attribution, we do need the capability for prediction in ML model, which is not trained in the design of current training approach.

Response:

Thank you for raising this important concern. In our study, we did not use the outputs from individual CMIP6 models as training samples. Instead, we used the multi-model ensemble mean surface air temperature as the target variable for the machine learning model. This approach eliminates the possibility of the model learning across different climate models.

In response to your concern regarding the strategy of random data selection for ML model training and testing, as suggested by another reviewer, we have revised our methodology by leaving one future scenario for model validation and other samples for model training. Specifically, we now hold out the entire SSP2-4.5 experiment as an independent test set, while using the remaining experiments—including 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—for model training, similar to the method recommended in the ClimateBench framework (Watson-Parris et al., 2022). This revised split ensures that the testing scenario is completely unseen during training, allowing for a more robust evaluation of the model's generalization and extrapolation capabilities.

After retraining the model with this new setup, we find that it continues to perform well on the SSP2-4.5 test scenario (Figures R1-R6), supporting its utility for attribution analysis. We have updated the manuscript accordingly to reflect these methodological improvements. We sincerely appreciate your thoughtful comments, which helped us enhance the rigor of our study.

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.

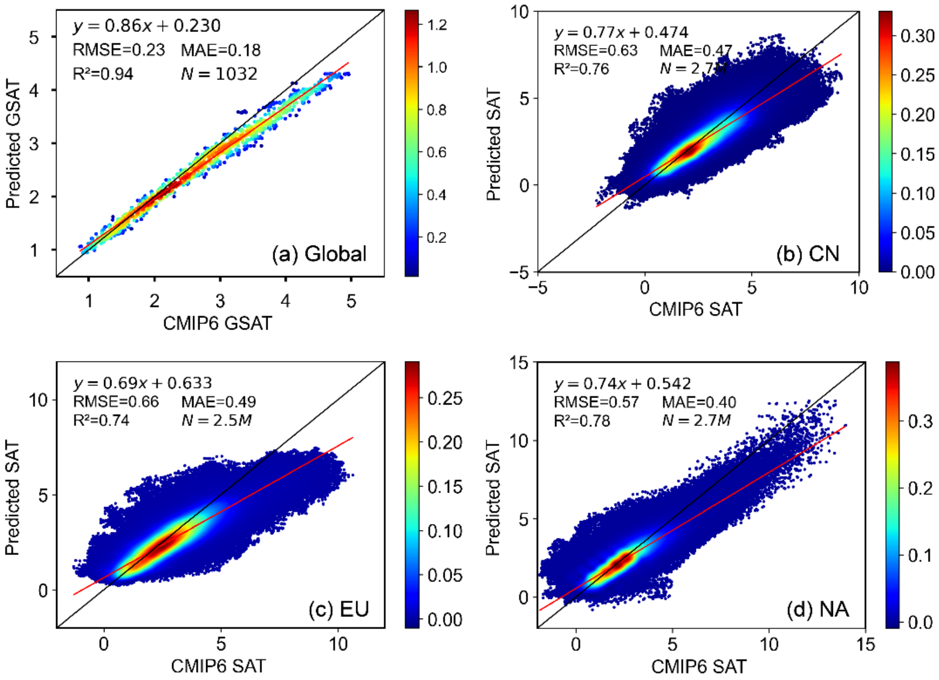

Figure R1. Scatterplot of the density of global and regional SAT (°C) over China, Europe, and North America from CMIP6 multimodel mean versus the predicted values from the LightGBM model under the SSP2-4.5 scenario. The black and red solid lines are the 1:1 lines and linear regression lines, respectively.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

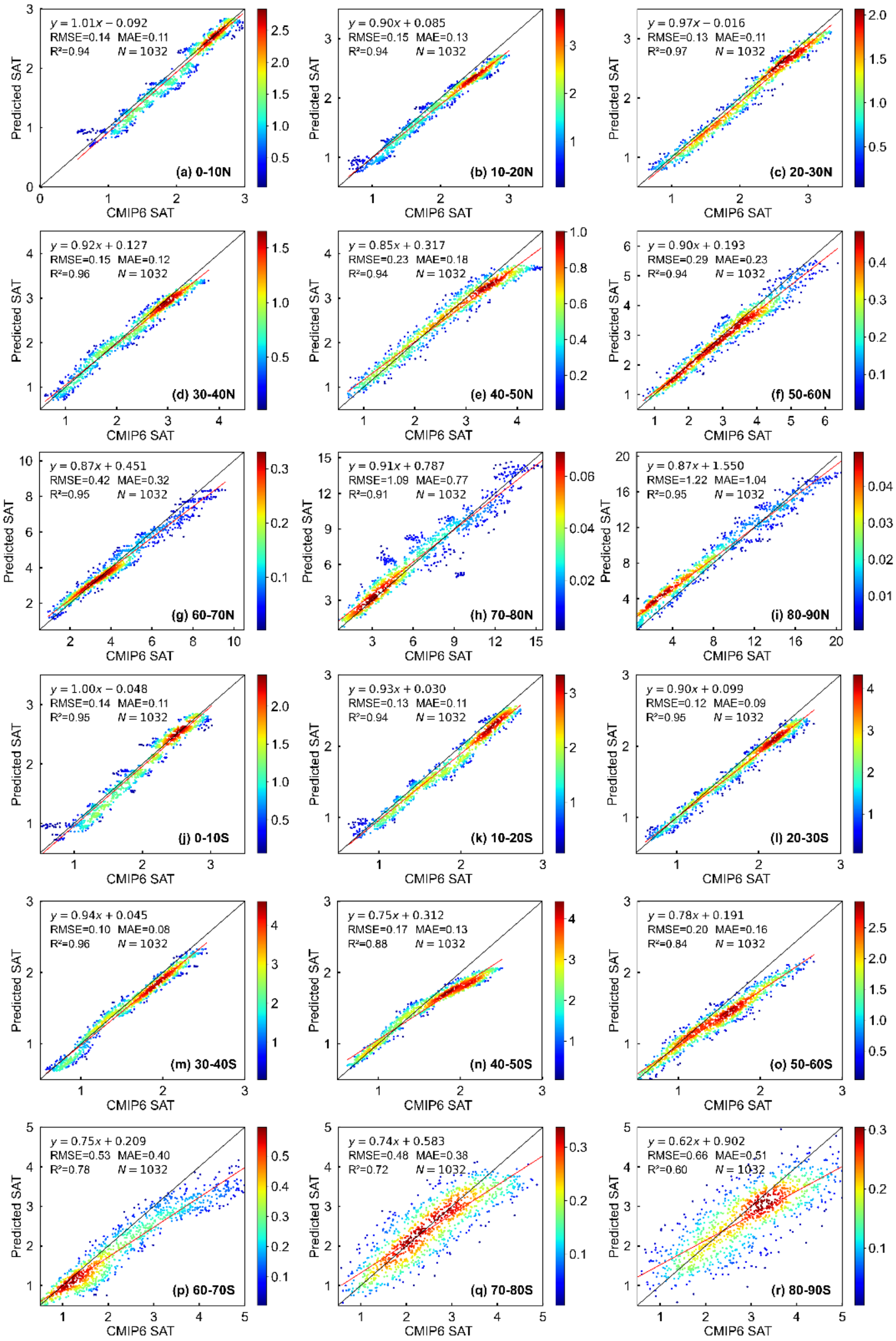

Figure R2. Scatterplot of the SAT density (°C) from CMIP6 multimodel simulations versus the predicted values from the LightGBM model for the eighteen latitudinal bands each spacing 10° from 90° S to 90° N, with color bars indicating the density of the data distribution. The black and red solid lines 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.

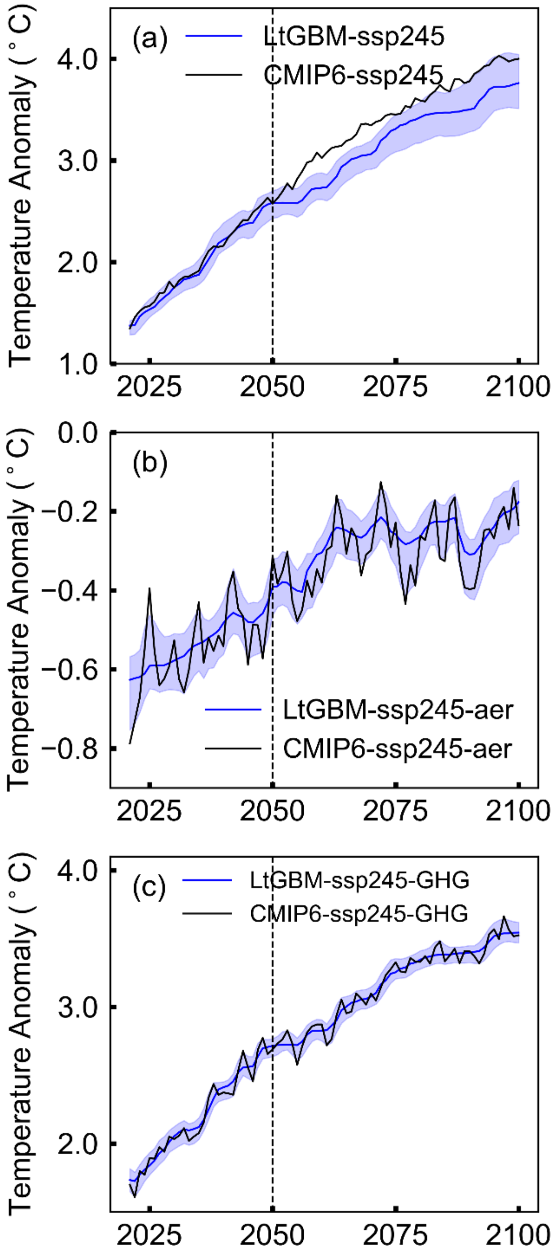

Figure R3. Time series of LightGBM-predicted GSAT anomalies (°C) and corresponding CMIP6 DAMIP values during 2021 − 2100 under the SSP2-4.5 scenario due to changes in (a) all forcing, (b) anthropogenic aerosols and (c) GHGs. Shaded areas indicate the range of the ML prediction by random separation of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

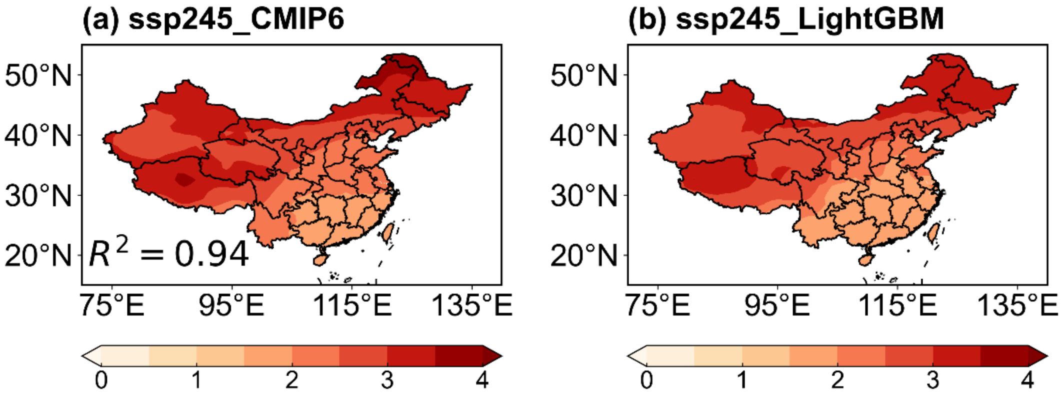

Figure R4. Spatial distribution of the 2015–2100 mean surface air temperature under the SSP2-4.5 scenario predicted by LightGBM and simulated by the CM IP6 multi-model ensemble.

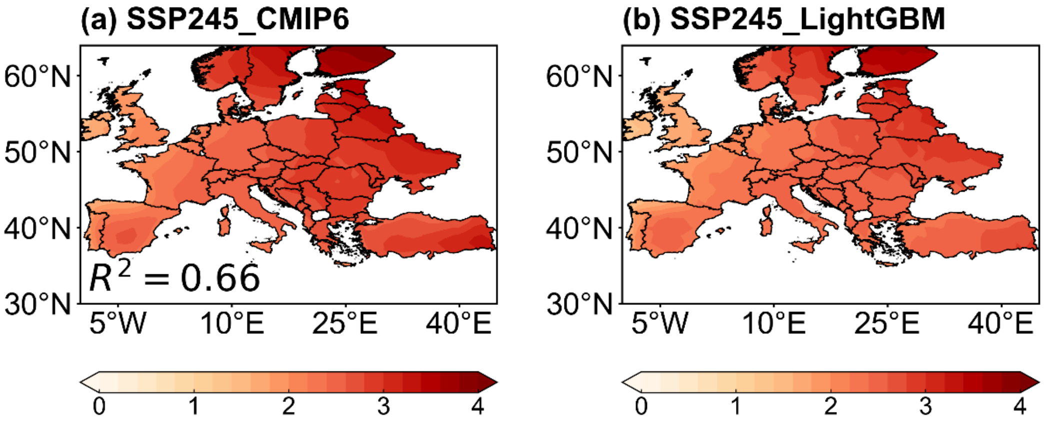

Figure R5. Same as Fig.R4, but for Europe.

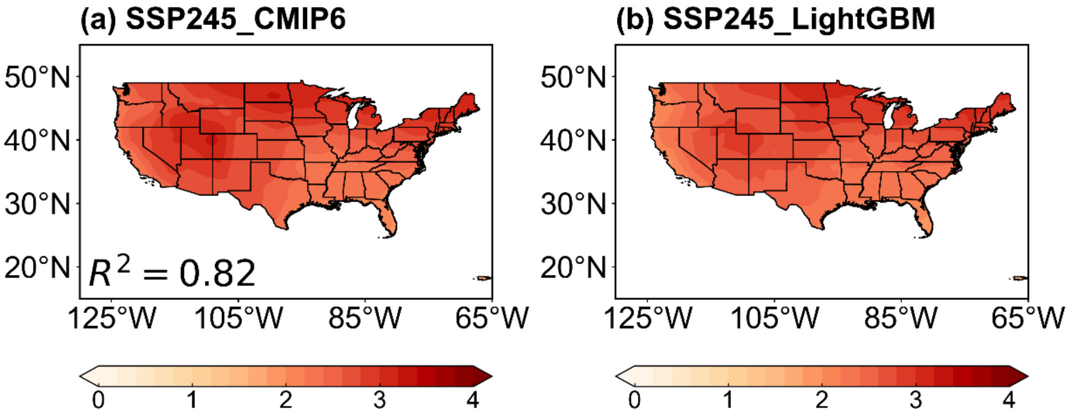

Figure R6. Same as Fig.R4, but for North America.

2. The description of ML is not clear enough in the method. Eg. What is import, what is output, how to train the ML etc.

Response:

We thank the reviewer for the valuable comment. We have not clarified the construction of the machine learning (ML) model in the Methods section. The input features of the model consist of anthropogenic forcing factors, including emissions of BC, OC, and SO_2 from anthropogenic and biomass burning sources, as well as concentrations of CO_2, CH_4, and N_2O, solar irradiance, volcanic forcing, and land-use changes. The model output is the surface air temperature averaged over global or regional scales for the corresponding time period. The revised manuscript has been updated accordingly to reflect these clarifications:

"In this study, the target variable for the ML models is the SAT anomaly, calculated as the multi-model mean from multiple CMIP6 experiments. Three types of SAT outputs are used as prediction targets: global mean SAT, zonal-mean SAT over 10^o latitude bands, and regional SAT for China, Europe, and North America. The experiments include historical simulations (referred to as "historical"), as well as single-forcing historical experiments from DAMIP: anthropogenic-aerosol-only (hist-aer), well-mixed greenhouse-gas-only (hist-GHG), natural-only (hist-nat), and CO_2-only (hist-CO_2). Future scenario simulations under SSP1-1.9, SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5 (hereafter referred to as ssp119, ssp126, ssp245, ssp370, and ssp585, respectively) were obtained from ScenarioMIP (O'Neill et al., 2016). In addition, single-forcing future simulations under SSP2-4.5 from DAMIP are used, including anthropogenic-aerosol-only (ssp245-aer) and well-mixed greenhouse-gas-only (ssp245-GHG) experiments. The DAMIP simulations adopted a "single-forcing" approach, where only one type of forcing varies while others are fixed at pre-industrial levels. Details of the experiments and the models employed are provided in Table S1.

To train the ML model, a range of temperature-related forcing variables is collected from the input4MIPs dataset, including aerosols and their precursors (BC, OC, and SO_2) emitted from anthropogenic and biomass burning sources, as well as GHGs concentrations, land use, solar radiation, and volcanic forcing. These data cover both the historical period and various future emission scenarios. A detailed description of the datasets is provided in Table 1."

"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^o S to 90^o 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."

3. What is the representative of the ML model? Does it represent a single CMIP6 model (which one?), or it represent the multiple model average?

Response:

We appreciate the reviewer's question. The machine learning (ML) model in our study is trained based on the multi-model mean surface air temperature derived from multiple CMIP6 models, rather than the output from any single climate model. Therefore, the predictions from the ML model represent the multi-model mean climate response rather than that of a specific model. This approach helps to reduce the influence of individual model biases and allows the ML model to more robustly capture the common relationships between external forcing factors and climate responses, which has been clarified in the manuscript.

4. As model-based attribution tech. pioneered by Prof Klaus Hasselmann, a key element for attribution is the inter-variability between climate models. Because this helps us understand the uncertainty and allows us to say that if the contribution of a climate forcer is significant enough to be detected, or not. I wonder how the multi-GCMs variability is represented in the ML model, and how is this been used to convince that the attributed aerosol/GHGs forcing is a significant fingerprint? Note that this multi-GCMs variability (stem from parameterization/structure/etc. uncertainties) is different from the shading area shown in Fig.5 (and many other figures), which only provides

the uncertainty of ML training.

Response:

Thank you for raising this important point. In our current modeling framework, we use multi-model mean surface air temperature outputs from many GCMs under the same experiment as the target variable for training the machine learning model. The input features, such as greenhouse gas concentrations and aerosol emissions, are derived from the input4MIPs project, which provides standardized external forcing datasets across all CMIP6 models. These inputs do not contain any information related to the internal structure, parameterization schemes, or physical processes of individual climate models. As a result, our machine learning model captures the relationship between external forcings and the multi-model mean temperature response, rather than responses from individual models.

This approach improves the signal to noise ratio and enhances the model's ability to identify the dominant climate forcing agents. It is suitable for assessing the relative importance of different forcings. However, we fully acknowledge that this methodology does not account for the variability in climate response arising from model structural uncertainties (Hasselmann, 1997). We recognize the importance of incorporating multi-model variability into machine learning based attribution studies and plan to explore strategies to integrate structural uncertainty into future work. This will help improve the robustness and comprehensiveness of the attribution results.

We have now added the corresponding discussion in the manuscript.

5. I think Fig.S4 is worth more interpretation. I cannot read the message (L236) from Fig.S4. In addition, could you please help me understand why in some cases that CO_2/CH_4/N_2O can contribute cooling in global and some regional scales (see Fig.S4 of the negative impacts on model temperature)?

Response:

Thank you for your valuable comment. Figure S4 presents the distribution of

SHAP values for each input feature, providing insights into both the importance and direction of each feature's contribution to the predicted temperature. The y-axis lists the input features, ranked by their overall importance in the model, with higher positions indicating greater influence. Each point represents one sample, and its color reflects the value of the corresponding feature (red indicates a higher value, blue a lower one). The x-axis shows the SHAP value, which represents the marginal impact of that feature on the model output. A positive SHAP value indicates a positive contribution to the predicted temperature, while a negative value indicates a negative contribution.

As noted, CO_2, CH_4, and N_2O exhibit negative SHAP values in certain samples. However, this does not imply that these greenhouse gases have a cooling effect in terms of physical climate processes. SHAP values are a relative, model-based measure that indicate how much each feature deviates a prediction from the model’s average output. In some cases, when the concentration of greenhouse gases is relatively low, while other warming agents such as black carbon or solar forcing are anomalously high, the model may attribute a negative marginal contribution to the greenhouse gases. Furthermore, due to the nonlinear nature of the LightGBM model and potential interactions or collinearity among features, it is possible to observe such negative SHAP values. This is a common phenomenon in interpretable machine learning and does not contradict physical expectations.

We have added the following description to the manuscript:

“Figure S4 presents the distribution of Shapley Additive Explanations (SHAP) values for each input feature, providing insights into both the importance and direction of each feature's contribution to the predicted temperature. The input features are ranked by their overall importance in the model and the color reflects the value of the corresponding feature. The SHAP value represents the marginal impact of that feature on the model output. A positive SHAP value indicates a positive contribution to the predicted temperature, while a negative value indicates a negative contribution. The SHAP analysis indicates that greenhouse gases, such as CO_2 and N_2O, have a positive contribution to the model-predicted temperature changes at both global and regional scales. Anthropogenic emissions of SO_2 also show a significant influence, primarily associated with decreases in global and regional SAT within the model predictions. Although BC aerosols are widely recognized for their warming effects, they do not exhibit the expected positive contribution in the machine learning model outputs. This discrepancy likely arises because sulfate aerosols dominate the historical climate forcing in the CMIP6 data, and the available CMIP6 simulations lack individual attribution experiments for scattering versus absorbing aerosols to serve as model inputs. These findings highlight the need for more comprehensive attribution simulations in future CMIP protocols to better capture the distinct climatic effects of different aerosol species."

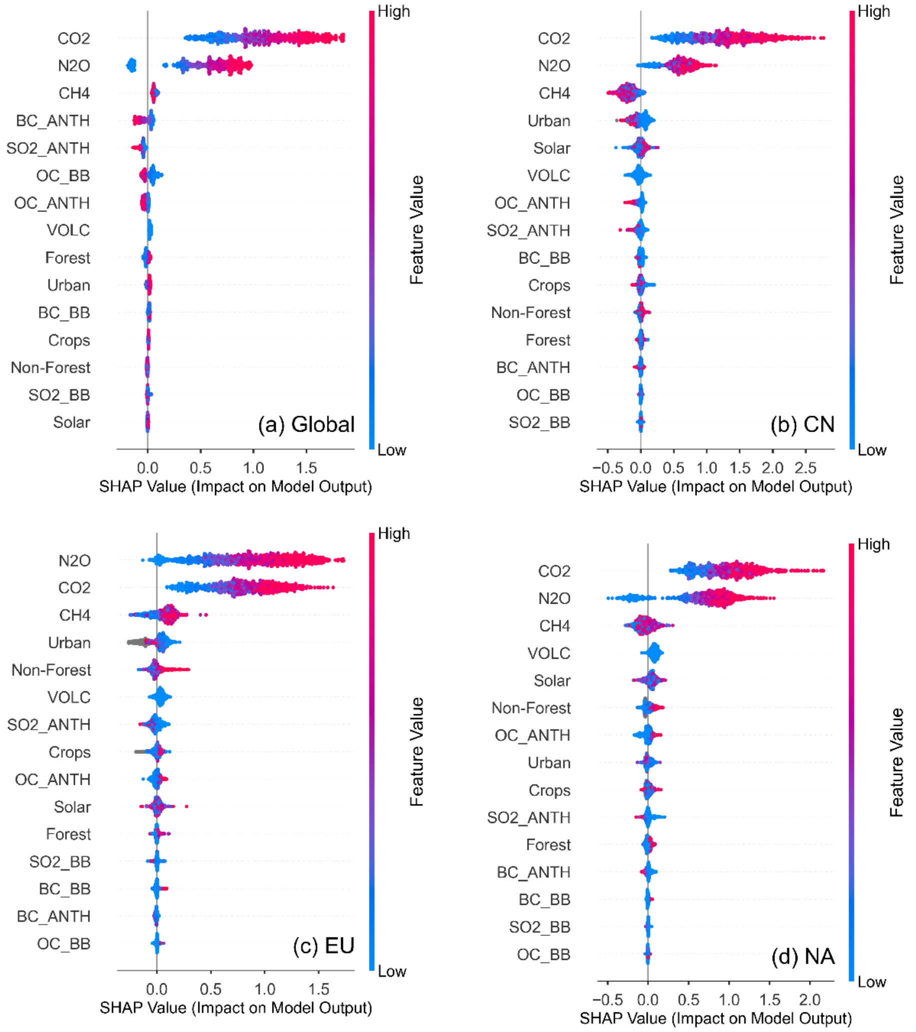

Figure S4. Shapley Additive Explanations (SHAP) values of LightGBM models for predicting (a) global and regional SAT over (b) China, (c) Europe, and (d) North America.

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.

Hasselmann, K.: Multi-pattern fingerprint method for detection and attribution of climate change, Clim. Dynam., 13, 601–611, 1997.