Reviewer 1

We appreciate the reviewer's insightful and constructive feedback, which has greatly enhanced the quality of the manuscript. We believe that the document has been significantly strengthened as a result of their thought-provoking questions and valuable remarks. We have responded to their key points as follows:

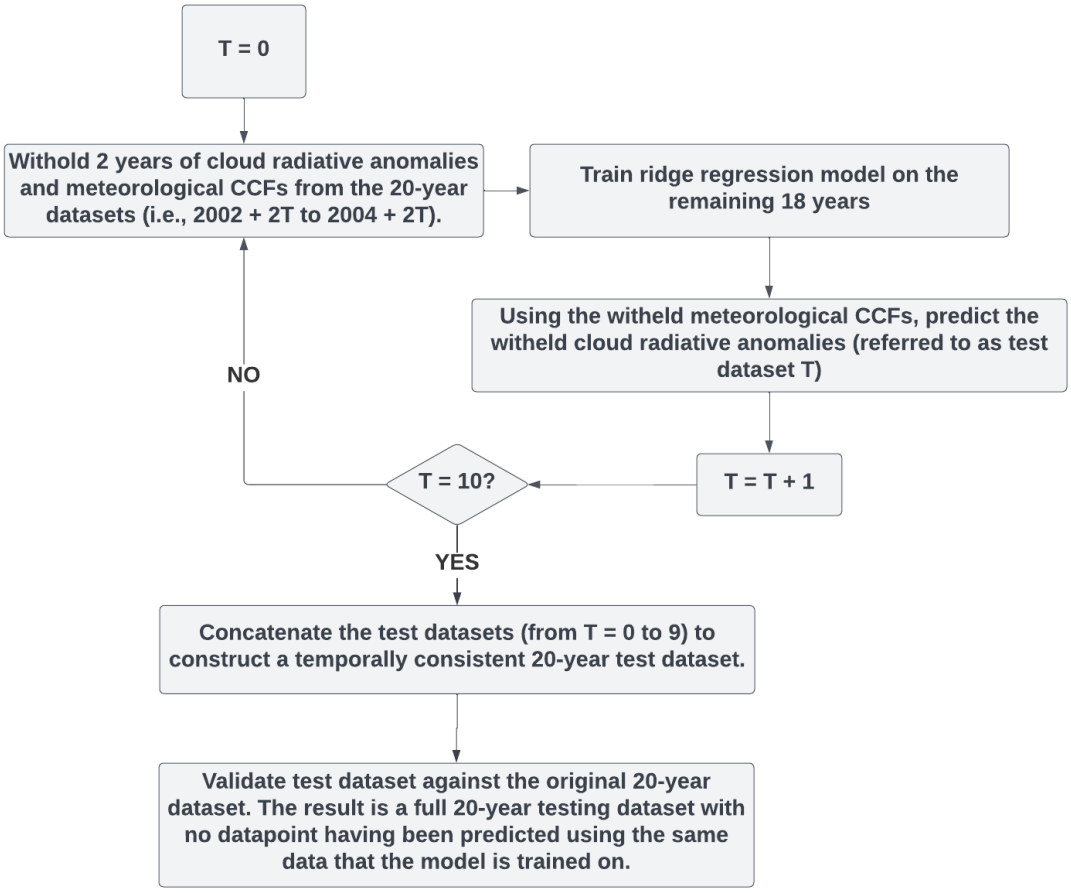
> *"In this paper, ridge regression is used to systematically evaluate the addition of five candidate cloud controlling factors (CCFs) to previously established core CCFs within large spatial domains to predict longwave high-cloud radiative anomalies. The results show that upper-tropospheric static stability is an important CCF for high clouds and longwave cloud feedback. All combinations of tested CCFs perform quite well for most locations at grid-cell scales, while differences between configurations for predicting globally-aggregated radiative anomalies are more pronounced. The authors found that spatial domain size is more important than the selection of CCFs for predicting local anomalies, and there is discrepancy between optimal domain sizes for local and globally-aggregated radiative anomalies.*
>
> *There are abundant technique details in the paper, and the method is potentially useful to evaluate the long-term high-cloud feedback. The paper might be accepted after addressing the following comments:*
>
> *Specific comments:*
>
> *1. In machine learning,* ***the dataset used to test the performance of a machine-learning model should be independent from the dataset that is used to train the data. What is the training dataset and testing dataset for the metrics of Fig. 2?*** *Ideally, the PI-control or AMIP simulations might be used as the training dataset, and abrupt4xco2 simulations might be used as the testing dataset.*

C1. We would like to reassure the reviewer that independent training and test datasets have been used throughout our analysis. We have used a rotating 2-year/18-year test/training dataset method to construct 20-years of predictions to compare against the full 20-year observed record. We have described our process with more clarity in the main text (line 296). We have also added a schematic (see Figure R1 below; this has been added to the supplementary, Fig. S2) that demonstrates our test/training process. This method is used consistently to determine the predictive skill shown in Figures 2 and S7 for both the observed and modelled data.

> *"For Sect. 5.1, 5.2 and 5.4 we use sensitivities to predict a two years validation dataset. We repeat this process, rotating the withheld data every two years resulting in ten unique training-validation dataset combinations (****see Supplementary Fig. S2 for a schematic of this process).***

We use historical observations and historical simulations to assess selections of cloud controlling factors instead of AMIP or Pi-control simulations. This is because sensitivities derived from historical, observed data have been used to constrain GCM simulated abrupt-4xCO2 cloud feedback (e.g., Ceppi and Nowack, 2021; Myers et al., 2021, etc., see main text for others). It is thus important that the sensitivities themselves are validated on the historical

time series. For the most analogous comparison between observations and atmosphere-ocean coupled models, we therefore also use historical CMIP simulations.



R1. Schematic demonstrating the process of rotating test-training datasets. This will be included in the supplementary material.

> *"For observations or historical simulations,* ***the first several or more years might be used to train the ridge regression model, and the last several years might be used to test the performance of the model****"*

C2. We have chosen to use our rotating test-training dataset method to reduce the sensitivity of our skill metrics to outliers. For example, if we test on only the last 2 years, and a single extreme is underestimated, the R^2 score is affected strongly by the outlier (owing to a small sample size of 24 datapoints). Our validation procedure in general allows us to more robustly estimate and compare the relative performance of the various controlling factor selections.

"2. R-square and r are highly relevant metrics, so I suggest using only one of them in the main text."

C3. We have expunged plots showing combined metrics so that Pearson r is the only metric shown in the main text. R^2 and RMSE are briefly discussed in Section 5.1 and shown

in the supplementary (Fig. S4/S5). This is to show that low R^2 doesn't necessarily mean low Pearson r or high RMSE.

*3. The **none-local effect of CCF on high cloud amount might be further explored and discussed**. The dependence of model performance to domain size might be associated with cloud transferring between adjacent grid boxes. In addition, previous studies suggest that the surface temperature in the tropics has significant impact on subtropical high cloud amount, is this process associated with the domain size dependence?*

C4. We have included further analysis and discussion regarding the non-local effect of CCF on high-cloud radiative anomalies. To summarise:

- Composite sensitivities have been analysed (with 2 shown in the supplementary Fig. S6 and shown below, and Figure R2). Composite sensitivities are constructed by domain-averaging independent local 21x11 sensitivities, where the target grid-cell matches some criteria (e.g., averaging all local sensitivities for Central Pacific tropical ascent grid-cells).
- Explained value fraction (EVF) has been assessed at 3 different domain sizes (also included in Fig. S6).

We have visually inspected composite sensitivities in conjunction with the EVF plots (**see R2 and Fig. S6 shown below**). Doing so, we find an emergent distinction between more "local" (such as UTRH, ω_{300} and RH_{700}) and more "non-local" (such as T_{sfc}, ΔU_{300} and S_{UT}) predictors. We find that our "local" predictors have decreasing EVF with domain size, whilst for non-local predictors the EVF increases with domain size. We also notice the non-local predictors have more spatially coherent patterns beyond the target grid-cell (e.g. wind shear in R2(b)), whereas for local predictors, the sensitivities appear noisier (much like "salt and pepper" noise) away from the target grid-cell, with largest magnitude concentrated close to the target grid-cell.

We propose that, while increasing domain size incorporates non-local contributions from T_{sfc}, S_{UT}, and ΔU_{300}, the larger domain adds potentially less relevant information for the more localised UTRH, ω_{300} and RH_{700} (also contributing a large proportion of the local EVF). The resultant predictive skill is thus a trade-off between additional relevant information (e.g., from T_{sfc}) and additional less-relevant information (e.g., from UTRH) which increases the domain size and thus the dimensions of the model. We have referenced the findings of this analysis in the bullet point list starting at line 412:

"

- *There is an emergent distinction between "local" and "non-local" predictors. For example, EVF for UTRH decreases with increasing domain size and, accordingly, we find that local UTRH sensitivities typically have strong magnitudes close to target grid-cell, with noisy, spatially incoherent coefficients further afield (see Fig. S6a-b for an example); thus, we describe UTRH as a "local" CCF (similarly for ω_{300} and RH_{700}).*

- *EVF for T_{sfc}, ΔU_{300}, and S_{UT} increases with domain size (i.e., "non-local" predictors), and each contribute a greater proportion of the globally-aggregated predictions compared to local predictions (Fig. S6c-d).*
- *Predictive skill is likely a trade-off between adding relevant information from "non-local" CCFs while adding superfluous information from "local" CCFs; i.e., too distant information does not provide additional predictive skill, at least to the degree that it would outweigh the corresponding increase in dimensionality of the regression problem.*
- *For globally-aggregated predictions, ω_{300} is the least important predictor (compared to the second most important for local predictions), thus explaining why the choice of pressure level of ω is less relevant at global scales (shown in Fig. 4) than locally.*

"

We have also included a new section in the supplementary material that discusses the discrepancies in domain size in more detail (Section S3), and we also agree that there are several mechanisms that may cause non-local dependence on the CCFs. For example, increased static stability to the east of a target grid-cell may be advected locally, or indeed an adjacent cloud transferring grid-cells. This is included explicitly in Section S3.

"Note that there are several mechanisms that may be associated with non-local sensitivities, including remote SST pattern effects for deep convection (Fueglistaler, 2019), the transferral of cloud from one grid-cell to another within the resolved time interval, or upstream/downstream advection of the meteorological drivers".

"Minor Comments:

It is recommended to check all instances of italicized text in the manuscript to ensure consistency throughout the text:

Line 99: delete the preposition "in".

Line 275: what is the variable "r"?

Line 287: where the second term on the right-hand side of Eq. (3) ...

Figure 1: The latitude and longitude coordinates should be marked on the map (and similarly for the subsequent figures).

Table 1: The formatting needs to be unified. For example, there are excessive gaps between certain words, and the sixth row of the table ("Key studies") lacks a space before it. Moreover, the last row ("Key studies") has a period, while the other rows do not."

Thank you for pointing out the above corrections, which we have implemented.

"Figure 9: I suggest adding an additional panel to compare the results with the CN21 method (i.e., CCFs containing only T_sfc, RH_700, UTRH, and ω_300). This comparison will better highlight the advantages of the new method."

C5. Thank you for your suggestion. We have broadened the discussion on Figure 9 to more directly compare against alternative configurations. Though we have chosen to not include an additional panel in Figure 9 showing CN (as visually, the spatial distribution of the radiative anomalies are similar between EIS and $S_{UT} + \Delta U_{300}$), we have assessed the absolute prediction error for a range of configurations during El Niño (where we calculate error = predicted anomaly – observed anomaly, and total tropical absolute error is the absolute sum of all tropical error). We look at the *absolute* anomaly to avoid rewarding configurations that produce compensating positive and negative prediction errors which cancel when averaging. We have also included an additional panel (panel (c)) showing the spatial distribution of the anomalies.

We find that EIS (ω_{500} – the CN configuration) has the highest absolute prediction error for R_{LW}, $R_{LW,CF}$ and $R_{LW,CTP}$. In fact, including EIS actually *increases* the absolute anomaly relative to just the core CCFs. Conversely, configuration S_{UT} (with no shear) has the lowest absolute error for $R_{LW,CF}$ and $R_{LW,CTP}$. For R_{LW} and $R_{LW,CTP}$, S_{UT} is followed by $S_{UT} + \Delta U_{300}$. These findings are mentioned in line 663:

"We also find that configurations S_{UT} and $S_{UT} + \Delta U_{300}$ predicts the tropical mean El Niño $R_{LW,CTP}$ with the smallest absolute error (not shown)".

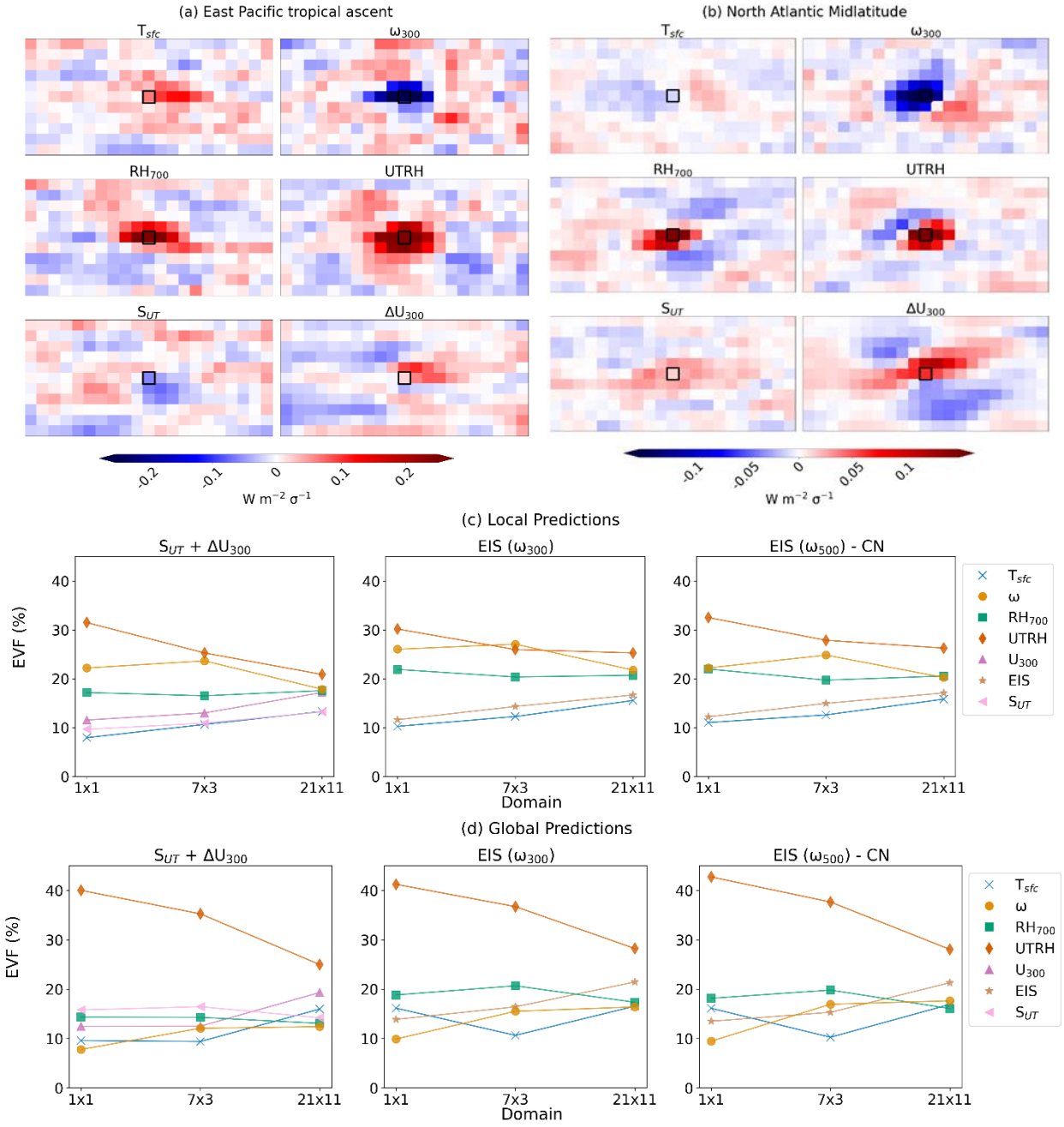
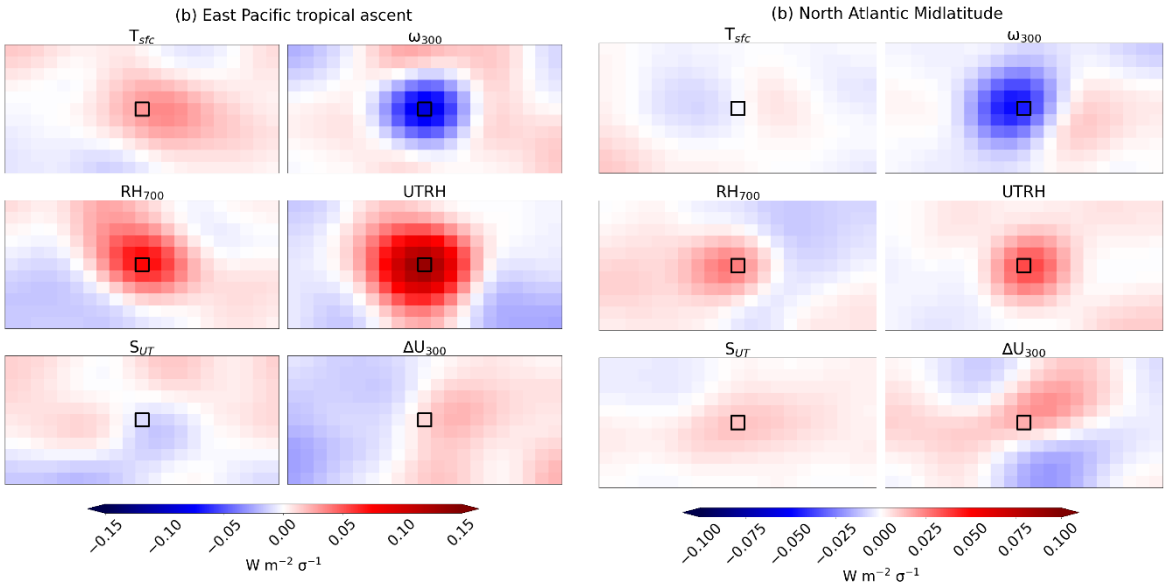


Figure S6. Composite spatial sensitivities using the 21x11 domain and configuration $S_{UT} + \Delta U_{300}$ (with additional core CCFs T_{sfc}, RH_{700}, UTRH, and ω_{300}) in (a) tropical ascent grid-cells (defined by climatological mean EIS < 1 K, and ω_{500} < 0 hPa s^{-1}) in the East Pacific (130°W to 80°W) and (b) North Atlantic (60°W to 10°E, latitudes north of 30°N) midlatitude clouds (climatological mean EIS > 1 K, and ω_{500} < 1.5x10^{-4} hPa s^{-1}). Panel (c) shows the global mean EVF as a function of cloud controlling factor and domain size for local predictions. Note that the global mean EVF has only been weighted based on latitude, and not as a function of R_{LW} standard deviation. Panel (d) shows the EVF for globally-aggregated predictions.



R2. Composite spatial sensitivities smoothed using a Gaussian filter (σ = 2) to reduce non-local noise. The 21x11 domain and configuration $\boldsymbol{S_{UT} + \Delta U_{300}}$ *has been used (with additional core CCFs* $\boldsymbol{T_{sfc}}$*, RH*$_{700}$*, UTRH, and* ω_{300}*) for (a) tropical ascent grid-cells (defined by climatological mean EIS < 1 K, and* ω_{500} *< 0 hPa s*$^{-1}$*) in the East Pacific (130°W to 80°W) and (b) North Atlantic (60°W to 10°E, latitudes north of 30°N) midlatitude clouds (climatological mean EIS > 1 K, and* ω_{500} *< 1.5x10*$^{-4}$ *hPa s*$^{-1}$*). These regions are the same shown in Figure S6.*

References

Ceppi, P. and Nowack, P.: Observational evidence that cloud feedback amplifies global warming, Proc. Natl. Acad. Sci. U.S.A., 118, e2026290118, https://doi.org/10.1073/pnas.2026290118, 2021.

Fueglistaler, S.: Observational Evidence for Two Modes of Coupling Between Sea Surface Temperatures, Tropospheric Temperature Profile, and Shortwave Cloud Radiative Effect in the Tropics, Geophysical Research Letters, 46, 9890–9898, https://doi.org/10.1029/2019GL083990, 2019.

Myers, T. A., Scott, R. C., Zelinka, M. D., Klein, S. A., Norris, J. R., and Caldwell, P. M.: Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity, Nat. Clim. Chang., 11, 501–507, https://doi.org/10.1038/s41558-021-01039-0, 2021.