Supporting information for "CERA: A Framework for Improved Generalization of Machine Learning Models to Changed Climates"

Shuchang Liu[1], **Paul A. O'Gorman**[1]

[1]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology

Corresponding author: Shuchang Liu, shuchang.liu@hotmail.com

Contents of this file

1 Training details

In each iteration, the autoencoder processes 8192 input samples, with 4096 from the control climate and 4096 from the +4 K climate. The reconstruction loss is computed on the full batch, and the latent alignment loss is evaluated between the control and +4 K samples. In the mean time, the downstream prediction loss is calculated using only labeled control-climate data. The final loss is a weighted sum of reconstruction loss, latent alignment loss and prediction loss to enable joint updates of the autoencoder and predictor.

The model is optimized using the AdamW optimizer with a learning rate of 3×10^{-3} and a weight decay of 10^{-3} for both the autoencoder and the MLP predictor. A learning rate scheduler with exponential decay and 4000 linear warm up steps is applied to both learning rates. Training is run for 30 epochs.

After initial tests on generalization performance, hyperparameters were selected through a sweep aimed at balancing latent alignment and predictive accuracy. The weight on the latent alignment loss, λ_{EMD}, was chosen as the largest value that did not lead to latent space collapse. When λ_{EMD} is too large, the encoder tends to collapse the inputs toward zero in order to minimize the EMD loss. We define latent space collapse as the average standard deviation of the latent variables falling below 0.1. Values of λ_{EMD} ranging from 10^{-1} to 10^{-5} were tested, and $\lambda_{\text{EMD}} = 10^{-4}$ was selected as the largest stable value. The weight on the supervised prediction loss, λ_{pred}, was progressively decreased from 1 to 0.001. Since $\mathcal{L}_{\text{total}}$ depends on this weight, we effectively selected λ_{pred} by minimizing the unweighted combination of the underlying loss terms. For λ_{pred} values between 1 and 0.01, $\mathcal{L}_{\text{pred}}$ varies by less than 5% at the end of training, and $\mathcal{L}_{\text{reconstruction}}$ remains small ($< 10^{-4}$). Interestingly, $\text{EMD}(Z_0, Z_{+4K})$ decreases when $\mathcal{L}_{\text{reconstruction}}$ becomes smaller, presumably because better reconstructions more strongly constrain the latent space. Consequently, $\text{EMD}(Z_0, Z_{+4K})$ reaches its minimum when λ_{pred} is set to 0.01. Further decreasing λ_{pred} below 0.01 led to an increase in $\mathcal{L}_{\text{pred}}$. Therefore, the best-performing configuration used $\lambda_{\text{pred}} = 0.01$. This combination provided a stable latent alignment without compromising predictive performance on the labeled control-climate data.

For the alternative version of our analysis (results shown in Fig. 3) which is used to enable direct comparison with Beucler et al. (2024), we fine-tuned the λ_{EMD} parameter to account for the modified input/output configuration. We found that we could use a larger λ_{EMD} to strengthen alignment, stopping before latent space collapse (defined as std < 0.1). We selected $\lambda_{\text{EMD}} = 10^{-3}$ as the largest value that preserved sufficient latent variability.

For the baseline models, we use the same architecture as CERA but without the autoencoder component. In particular, the predictor network has identical depth and width, and we employ the same optimizer, training schedule, learning rate (3×10^{-3}), and weight decay (10^{-3}). We tested a range of the aforementioned shared hyperparameters and found the models to be largely insensitive within reasonable bounds. Accord-

ingly, we used the same hyperparameter values across models to ensure a fair comparison.

2 Sensitivity experiments

2.1 Kernel size

We performed a sensitivity experiment for using a kernel size of 3 (kernel3) which implies that the autoencode becomes non-local in the vertical. No padding was applied, resulting in a reduction of one grid level from each boundary (two levels total) per layer. The latent dimension decreased to 3 (channels) $\times$ 14 (levels). In this configuration, we needed to decrease λ_{EMD} and λ_{pred}, because the latent space collapsed and the reconstruction loss increases compared to the kernel size of 1. The results reported here use $\lambda_{\text{EMD}} = 3 \times 10^{-5}$ and $\lambda_{\text{pred}} = 0.005$. The resulting performance is somewhat worse than CERA with kernel size 1 (Fig. S2).

Further increasing the kernel size results in excessive information loss within the latent space. Because the latent space is constrained to three channels, the reduction in vertical levels caused by larger kernels prevents the model from retaining sufficient structural detail. We also experimented with applying layer normalization across the vertical dimension to enhance vertical communication, but did not observe performance improvements.

2.2 Aligned channels in the latent space

CERA was designed to align only two of the three channels in the latent space. To validate this choice, we conducted an experiment in which all latent dimensions were aligned and used for prediction of the outputs (CERA_all_channels; Fig. S4). The results indicate that the configuration with one unaligned channel leads to a small improvement over the fully aligned version.

2.3 Hyperparameters

We have added sensitivity tests for λ_{pred} (Fig. S5) and λ_{EMD} (Fig. S6). Varying λ_{pred} from 0.01 (CERA) to 0.05 (pred_0.05) and 0.005 (pred_0.005), yields comparable R^2 values. Similarly, changing λ_{EMD} from 1×10^{-4} to 4×10^{-4} (EMD_4e-4) and 2×10^{-5} (EMD_2e-5) results in comparable performance. However, when λ_{EMD} is increased to 8×10^{-4} (EMD_8e-4), model peformance degrades substantially, likely due to the collapse of the latent space for some random seeds.

2.4 Training CERA on both control and warm climate

We conducted an additional experiment in which we trained CERA and the Baseline model on both control and +4K climate inputs and outputs (Fig. S7). This expanded training set contains twice the volume of the original control-only data for the predictor. While training on both climates improves general performance as expected, CERA_bothClimate remains superior to Baseline_bothClimate. This suggests that the autoencoder and alignment are beneficial even when warm climate output data are available.

3 Formula for instantaneous surface precipitation rate

The instantaneous surface precipitation rate for both the ML models and the high-resolution simulation is computed for offline results by vertically integrating the micro-

physical tendency of total condensate, q_{T_mic}, with density weighting:

$$P_{\mathrm{tot}}(z=0) = -\int_0^\infty \rho_0 q_{T\_\mathrm{mic}} dz. \tag{1}$$

For simplicity, we exclude the surface ice sedimentation flux, which is typically small. This formulation is similar to Equation S6 of Yuval et al. (2021), but we note that their equation omits a negative sign.

4 Extension to hemispherically asymmetric SST boundary conditions

To test CERA under broader conditions, we employed simulations featuring a hemispherically asymmetric SST configuration (perpetual solstice and an off-equatorial SST maximum), which substantially alters the global circulation. A detailed description can be found in Text S3 in Yuval and O'Gorman (2023).

We use CERA in this case to improve generalization from a hemispherically symmetric climate to a hemispherically asymmetric climate. We use inputs from both symmetric (control) and asymmetric simulations, while restricting the target outputs to the symmetric simulations. Results in Fig. S8 indicate that this task is less challenging than generalizing to the +4K climate, showing less severe performance degradation for the Baseline MLP. CERA performs better than both Baseline and RH+B Baseline in both the control climate and the asymmetric climate. It is interesting that RH+B baseline does not outperform the standard baseline in this context, which might be because the RH+B transformation is specifically tailored for generalizing to a warmer climate.

5 Latent variable distributions and Earth Mover's Distance for different models

We visualized the distribution of the inputs and latent space to show how Earth Mover's Distance (EMD) is reduced using CERA (Fig. S9-S12). To make the distributions comparable across experiments and levels, we normalized the inputs and latent space at each vertical level by the mean and standard deviation of the control climate data.

We calculated the EMD for normalized inputs and latent space, averaging the results across all inputs for the Baseline experiment and aligned channels across all five seeds for CERA and CERA-noAlign experiments. The overall EMD loss indicates that CERA achieves the most aligned distribution (EMD=0.77), while CERA-noAlign (EMD=1.02) already demonstrates strong alignment compared to the raw inputs (EMD=1.75). Note that these numbers are sensitive to the choice of distance and how the data is normalized which can affect which parts of the distributions or which vertical levels are emphasized.

References

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., ... O'Gorman, P. A. (2024). Climate-invariant machine learning. *Science Advances*, *10*(6), eadj7250.

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophysical Research Letters*, *48*(6), e2020GL091363.

Yuval, J., & O'Gorman, P. A. (2023). Neural-network parameterization of subgrid momentum transport in the atmosphere. *Journal of Advances in Modeling Earth Systems*, *15*(4), e2023MS003606.

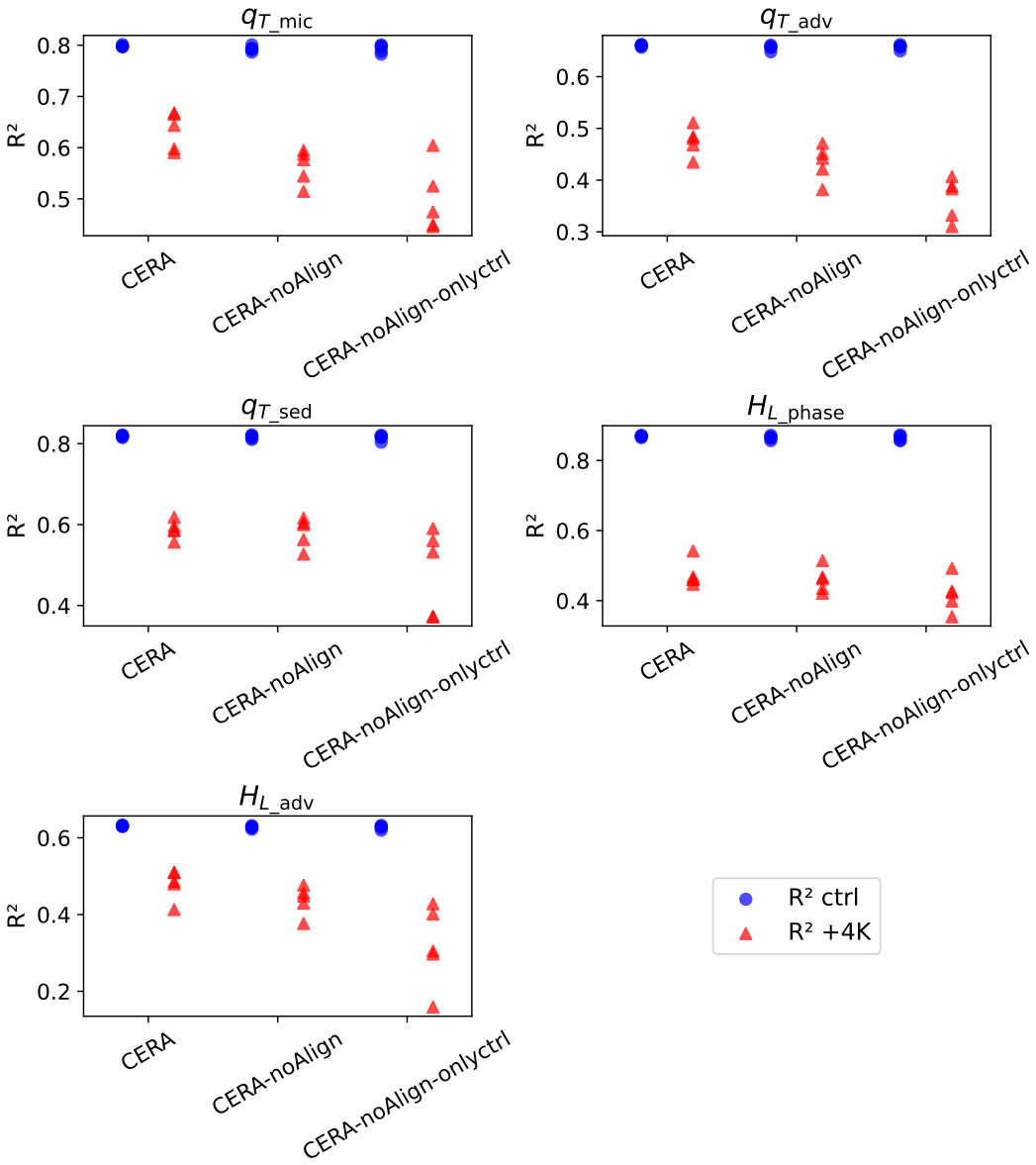

Figure S1. Similar to Figure 2, but with the addition of a version of CERA-noAlign that is trained on control data only (CERA-noAlign-onlyctrl).

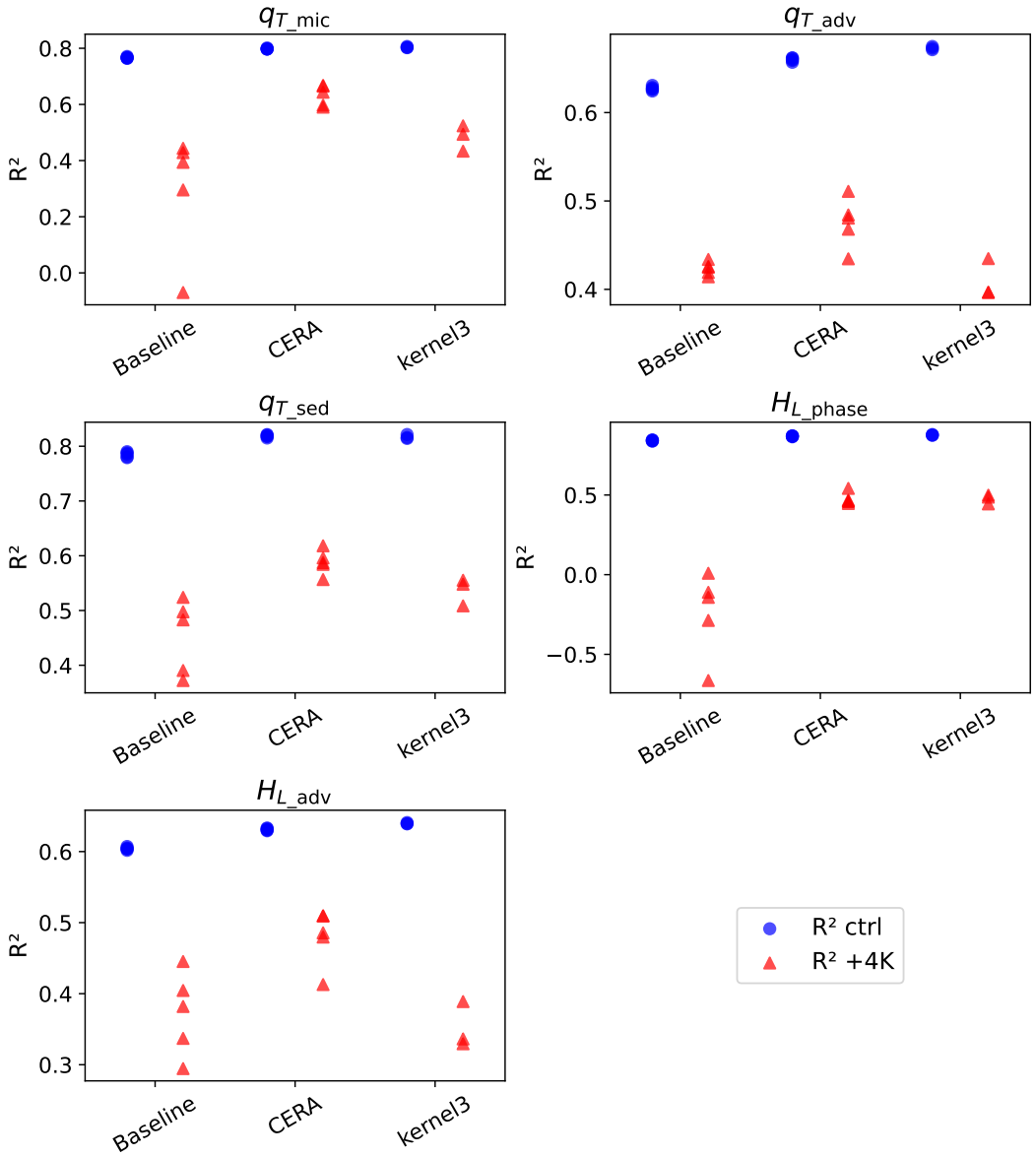

Figure S2. Similar to Figure 2, with the addition of results for a kernel size of 3 (kernel3).

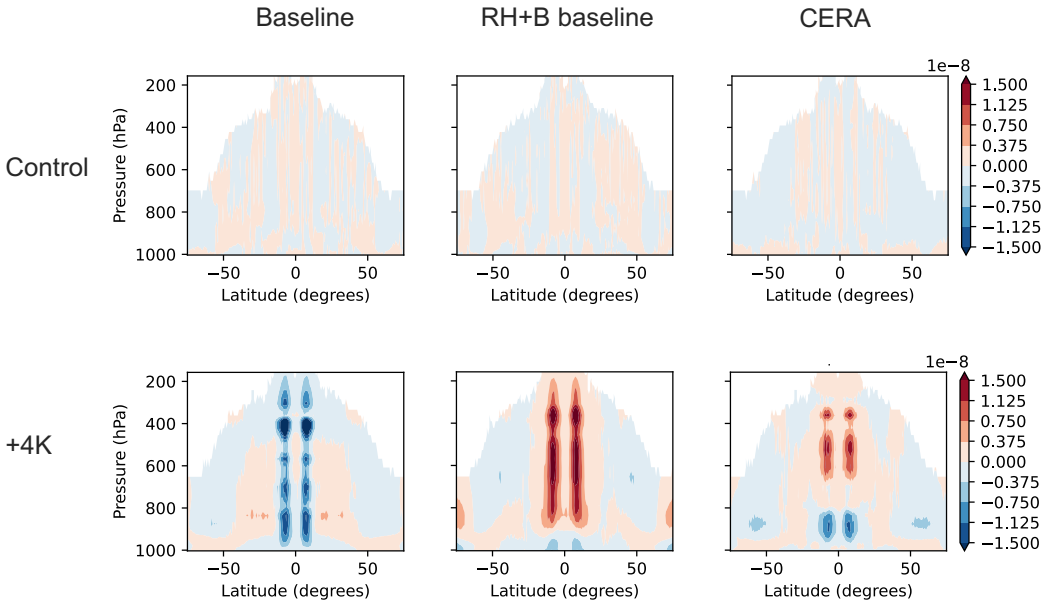

Figure S3. Offline error for q_{T_mic} versus latitude and pressure. Shown are the coefficient of determination (R^2) values for predicted q_{T_mic} (moisture tendency due to microphysical conversion to precipitating water) in the control and +4 K climates, averaged over five random seeds.

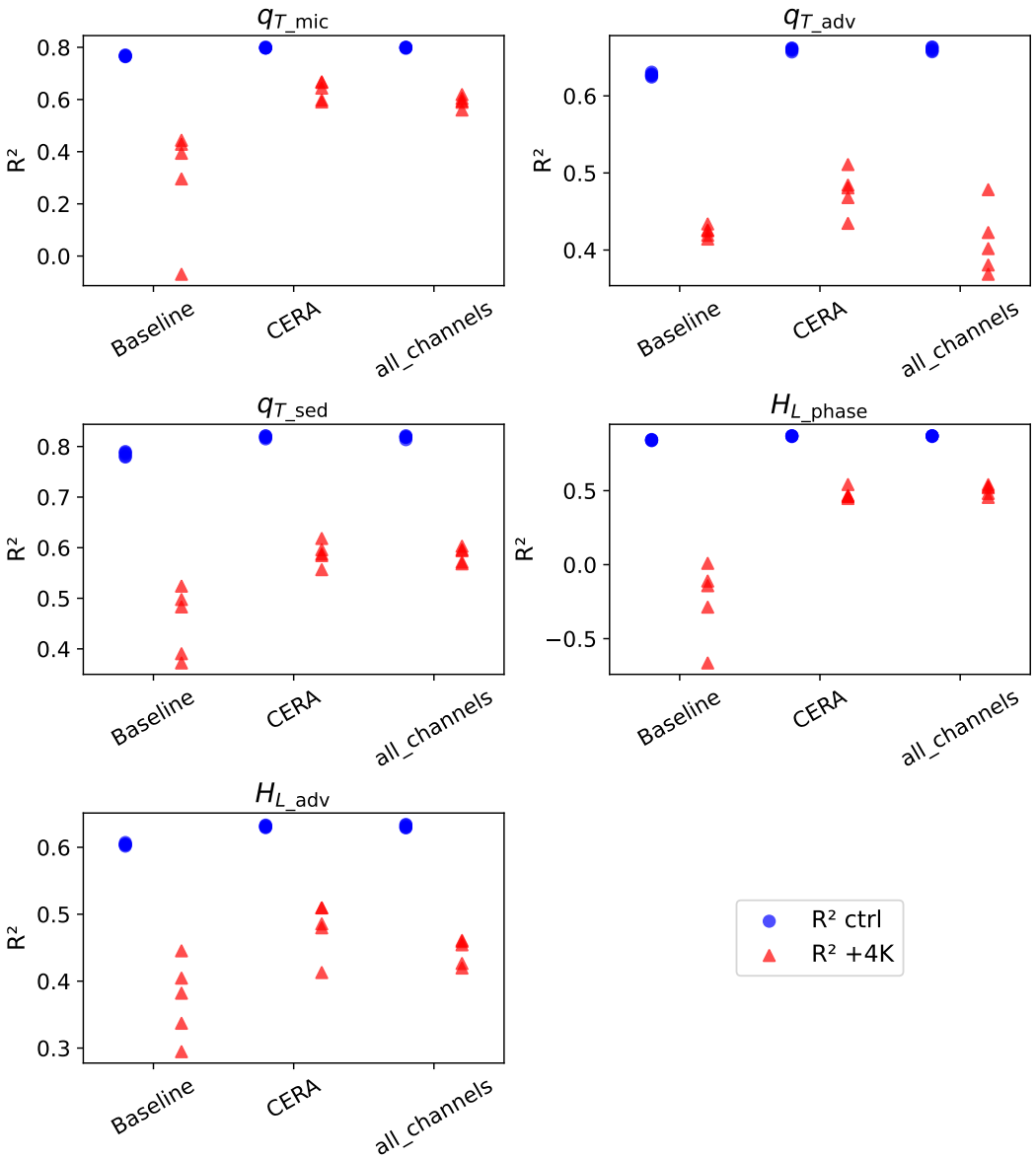

Figure S4. Similar to Figure 2, with the addition of results for the experiment where all latent dimensions were aligned (CERA_all_channels).

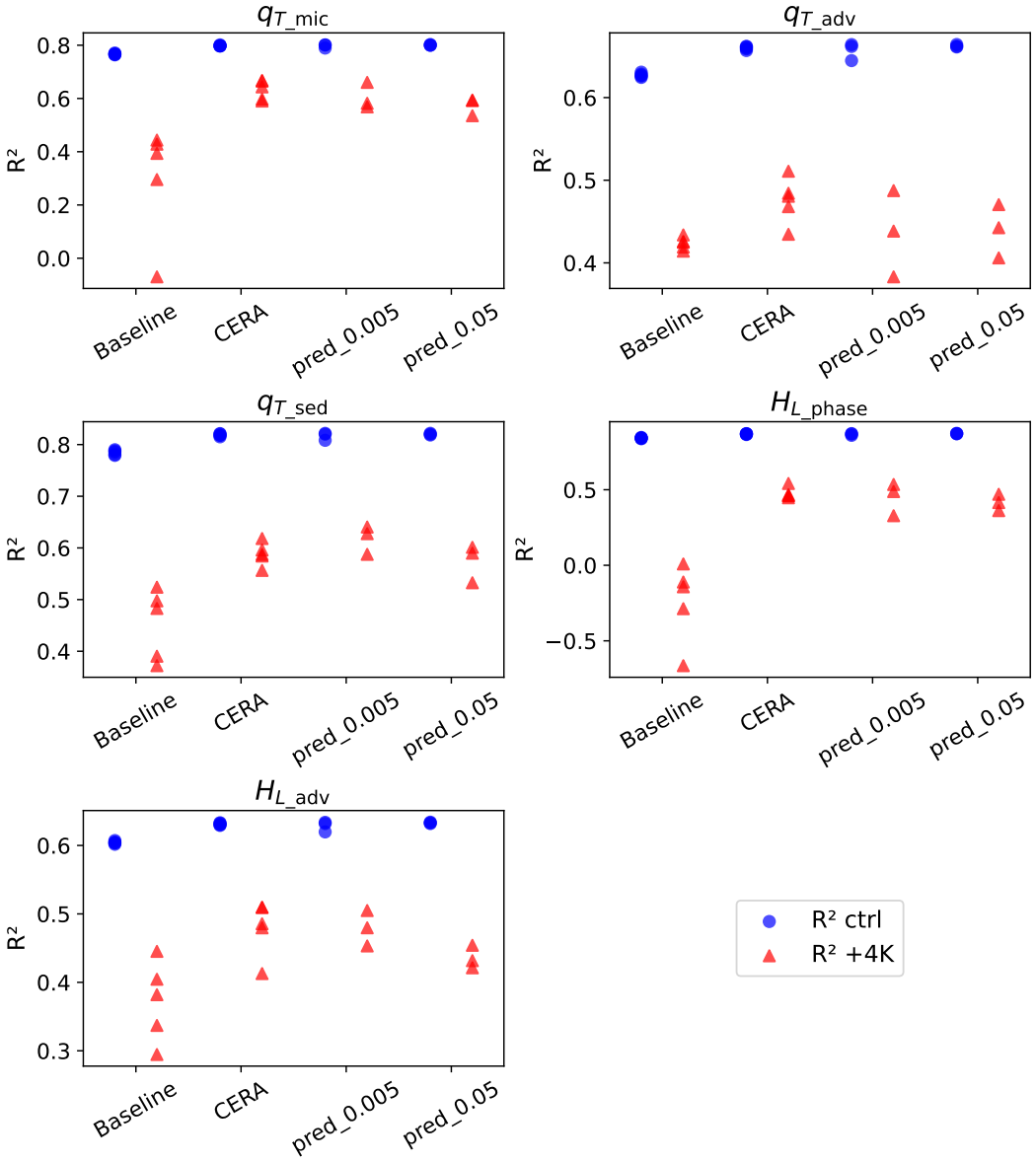

Figure S5. Similar to Figure 2, with the addition of sensitivity experiments regarding λ_{pred} used in CERA training. Three individual random seeds were used in training for the sensitivity tests.

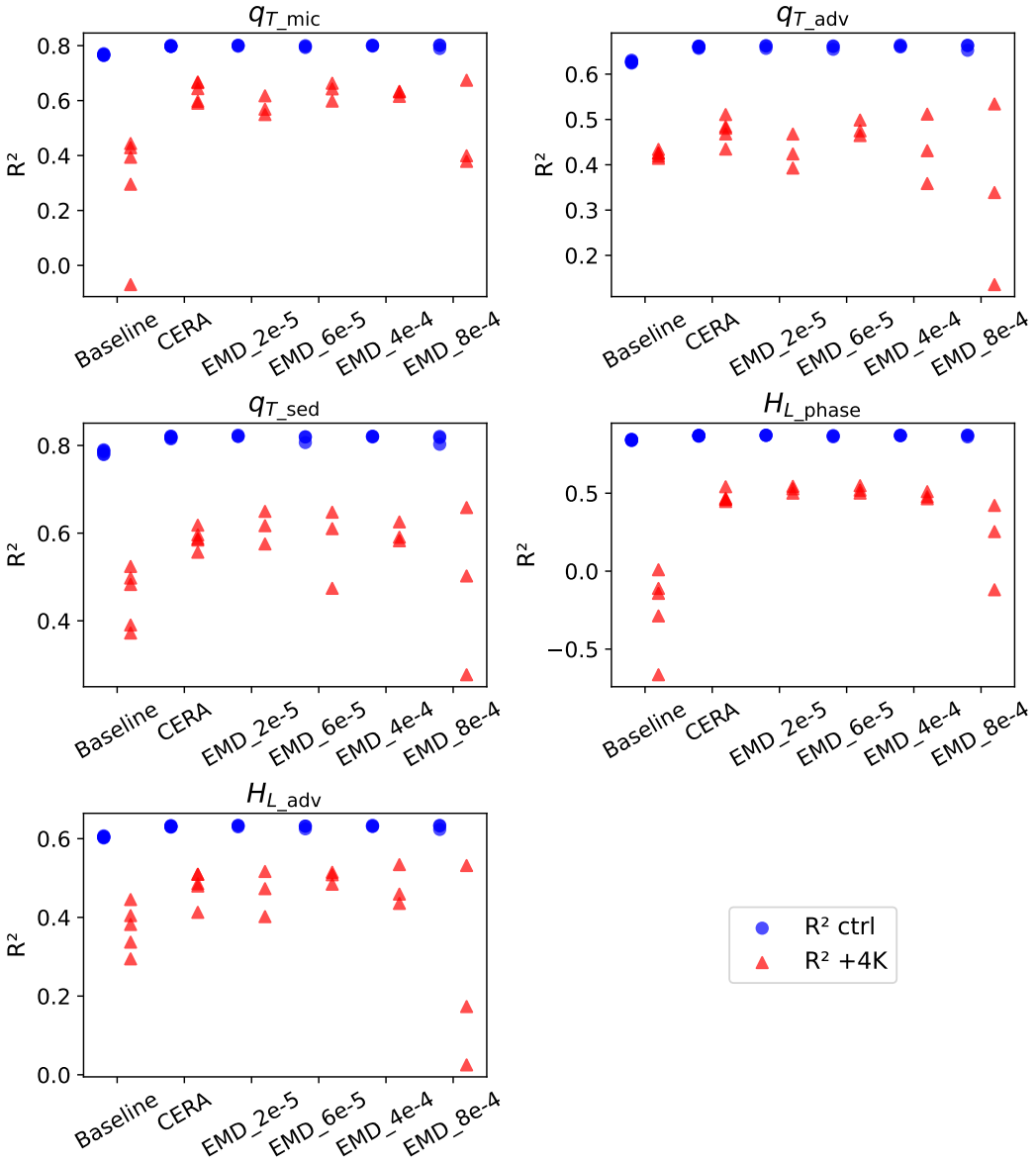

Figure S6. Similar to Figure 2, with the addition of sensitivity experiments regarding λ_{EMD} used in CERA training. Three individual random seeds were used in training for the sensitivity tests.

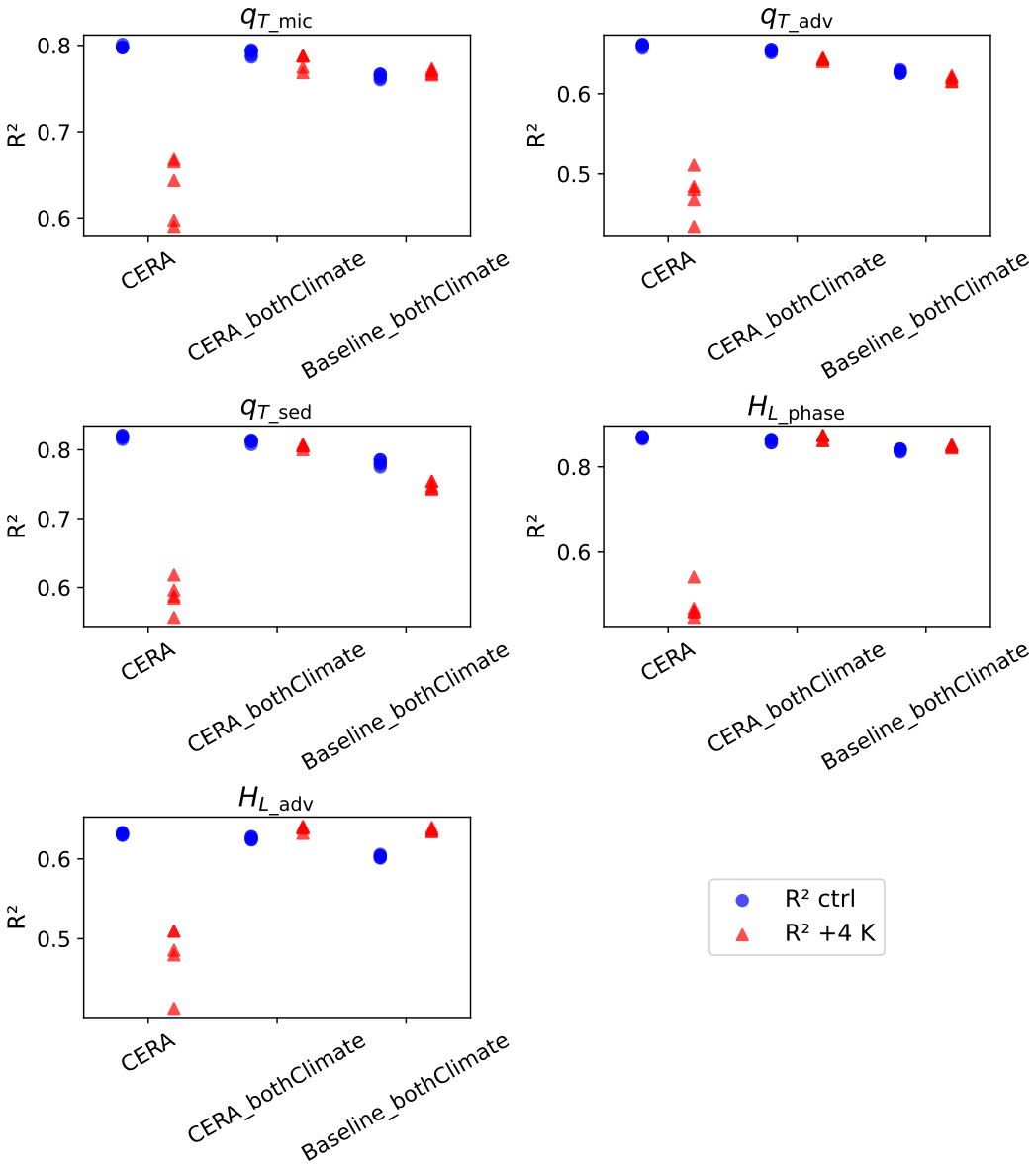

Figure S7. Similar to figure 2, but showing results when CERA and Baseline MLP are trained on inputs and outputs from both climates.

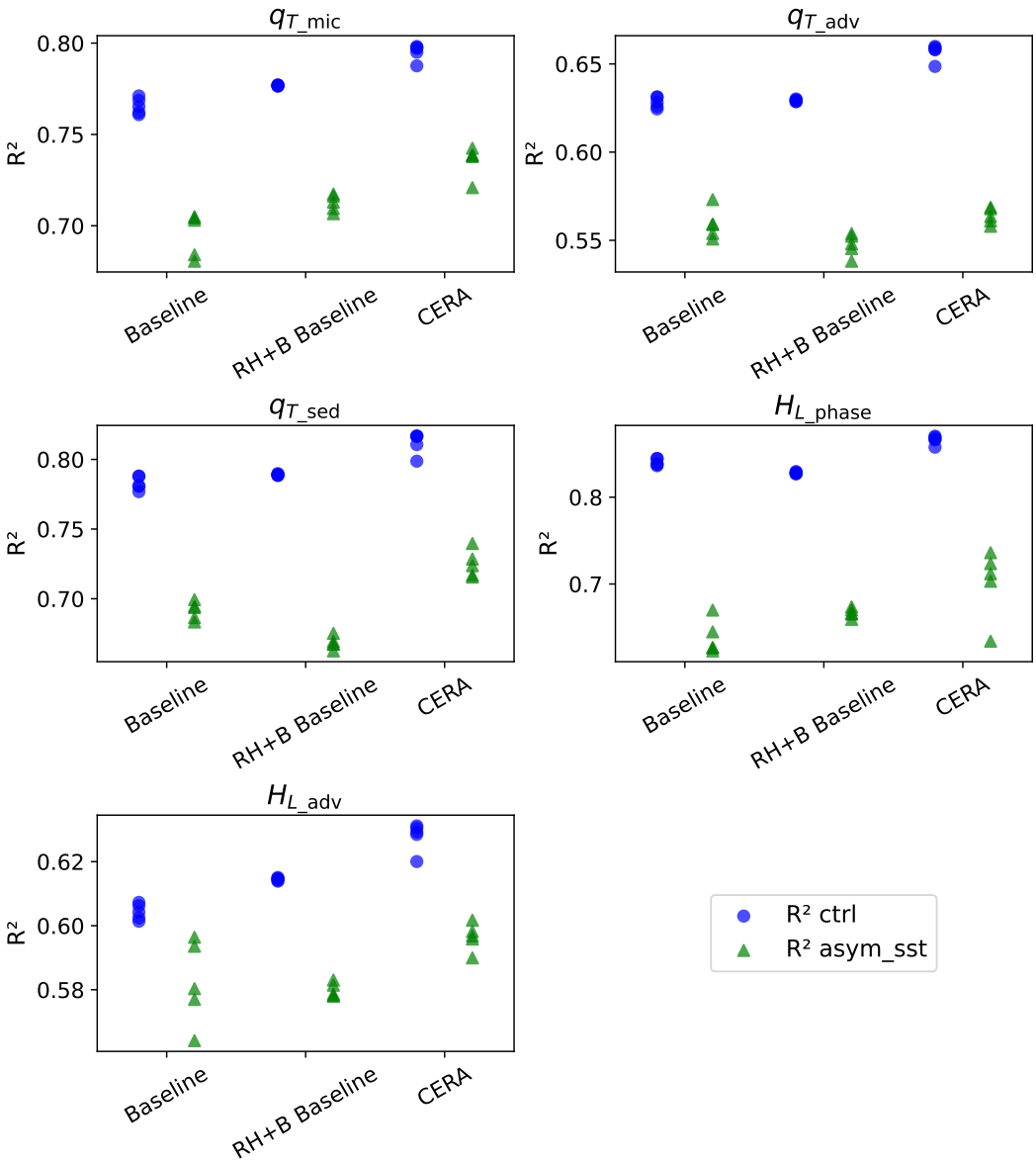

Figure S8. Similar to Figure 2, but for generalizing from the control (hemispherically symmetric) climate to a hemispherically asymmetric climate.

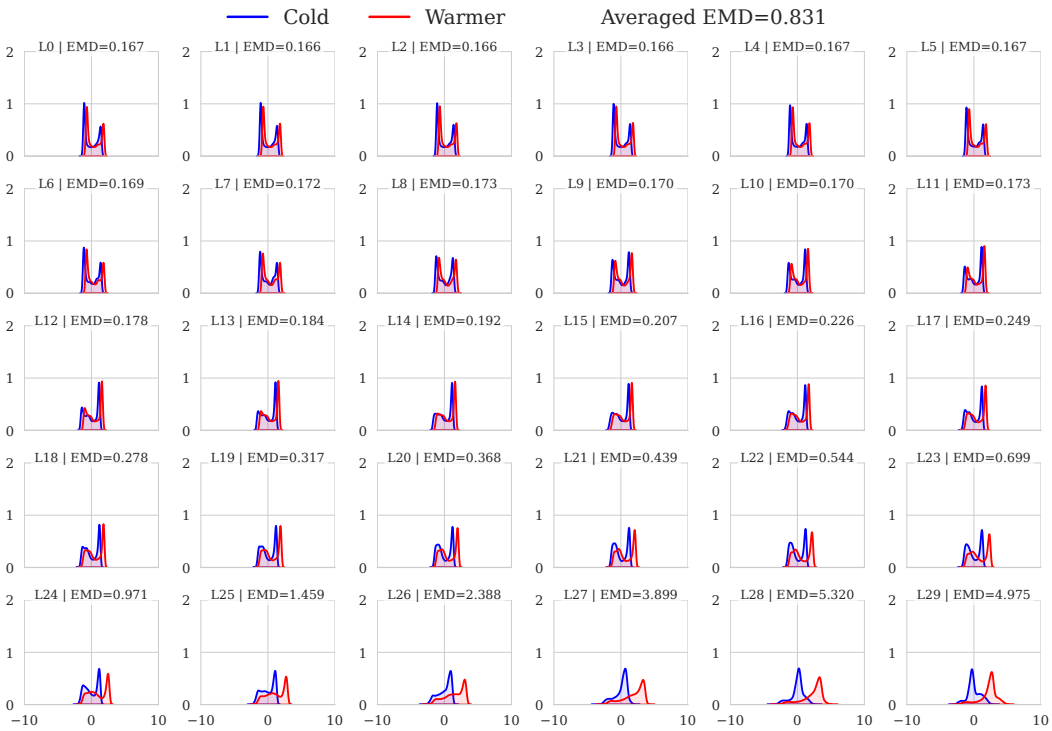

Figure S9. Distribution of the input temperature at each vertical level.

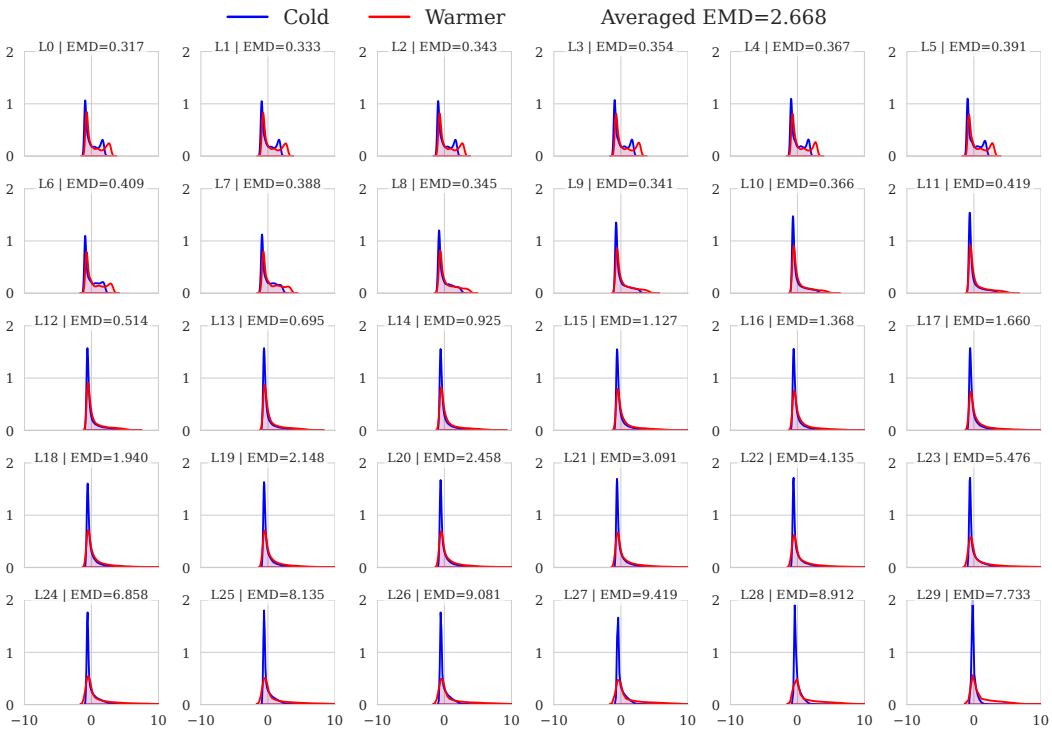

Figure S10. Distribution of the input specific humidity at each vertical level.

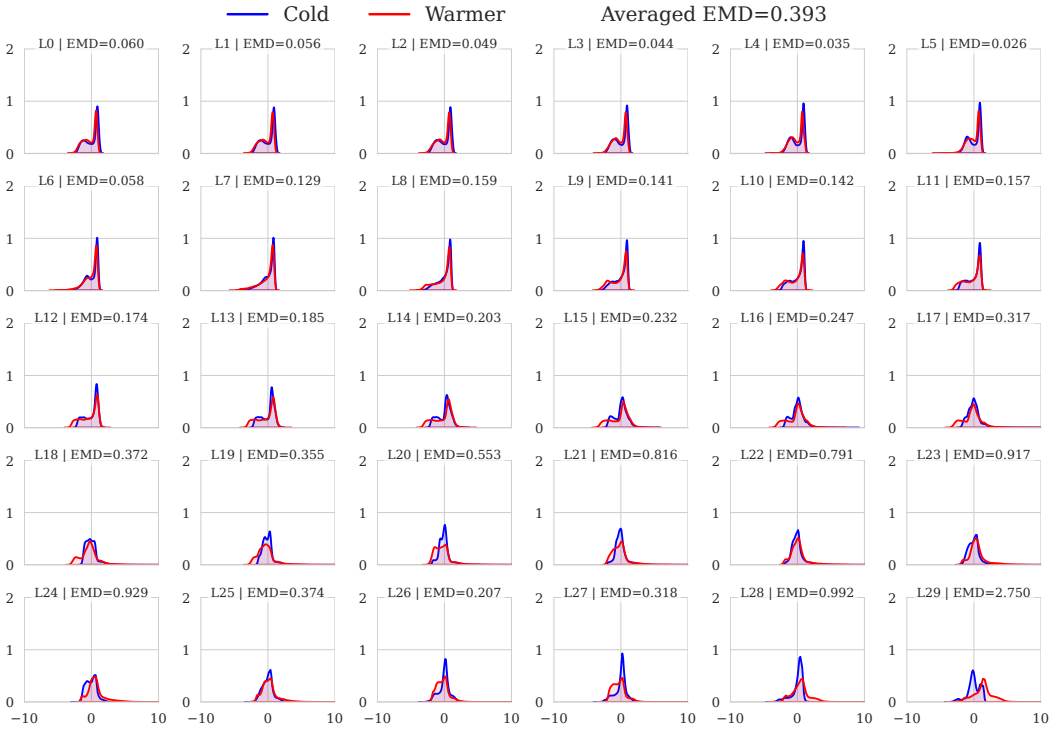

Figure S11. Distribution of the first channel of the latent space of CERA at each vertical level. One random seed was randomly chosen for the plot.

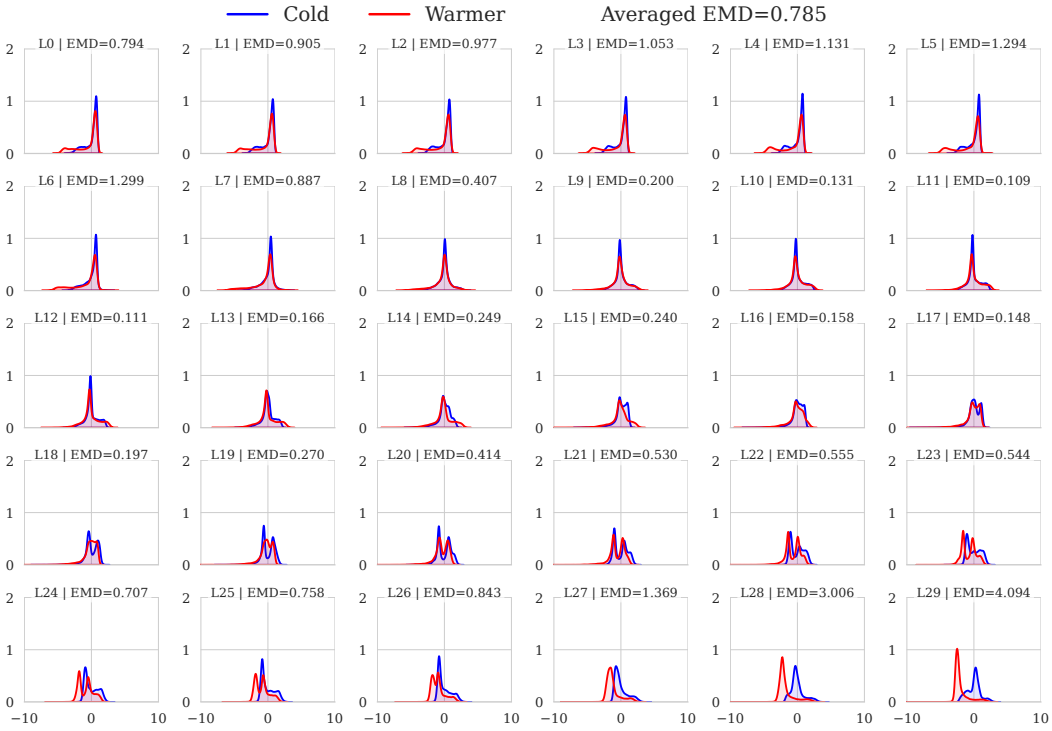

Figure S12. Distribution of the second channel of the latent space of CERA at each vertical level. Same random seed as in Figure S11.