Major Comments:

1. In Figure 3, it appears that LUCIE-3d has a temperature bias of 0.5 K in its mean state. LUCIE-3d and LUCIE-3d SST have the right temperature trends, but the absolute global mean temperature is off by ~0.5-0.7 K for the surface. Why does this happen? This problem also exists in GCMs in AMIP, but in ML models, wouldn't the expectation be that this bias does not exist since they are trained on ERA5 itself? NeuralGCM (https://research.google/blog/fast-accurate-climate-modeling-with-neuralgcm/) and ACE2 do not appear to have this mean state bias.

Author's response: We thank the reviewer for this comment. We are aware of this problem in LUCIE-3D. We believe that the source of this problem lies in the absence of TOA (top-of-the-atmosphere) radiation variable in the diagnosed quantities and thus an absence of an energy budget in the loss function. We do not see an immediate connection to the similar consequence in GCMs. In future iterations of this model, we are incorporating TOA as a variable as well and we believe that will address this specific problem of an underestimated global mean T2m.

2. Pg. 7 Line 138: Why does the model prescribe land temperatures at 270 K? Land temperatures have a significant diurnal cycle and variability. Prescribing them at 270 K isn't the correct value, and it seems like it would limit the ability to use LUCIE-3D for downstream applications. Either land temperatures should be prognosed (as in ACE2) or they should not be included at all (NeuralGCM), but why set them at 270 K? It also seems that this choice makes the model more brittle. It is worrying that the authors have to apply additional smoothing in order to get a reasonable response to +2K SST, and that without the smoothing, the model's land response is of the wrong sign. (While other emulators have errors with +2K SST, they were not fundamentally of incorrect sign over land for temperature and moisture).

Author's response: We thank the reviewer for this question. LUCIE takes atmospheric variables including near surface temperature which is directly correlated to the land temperature. The fixed unrealistic value over the land in SST is simply a mask for the model to recognize and ignore the land when processing sea surface temperature. Since SST is only used as a forcing variable, fixing the land value is not proposed to help the model recover the land temperature but to enlighten the model with SST as a driving force of the atmosphere. We have experimented with mapping the near surface temperature to the land values of the SST and update the values with prediction beside this smoothing technique. We simply observed a better performance with the current setup. +2 K SST is an open research topic and to our best knowledge, ACE2 is indeed showing the wrong sign especially over the land with +2 K and +4 K SST. The current setup is a

proposed potential solution and our effort on solving this problem and we indeed show accurate signs in the responses for both +2K and +4K SST perturbation. In Fig. 1 (of this document) we show a comparison of responses of ACE2 and LUCIE-3D for different SST perturbations.

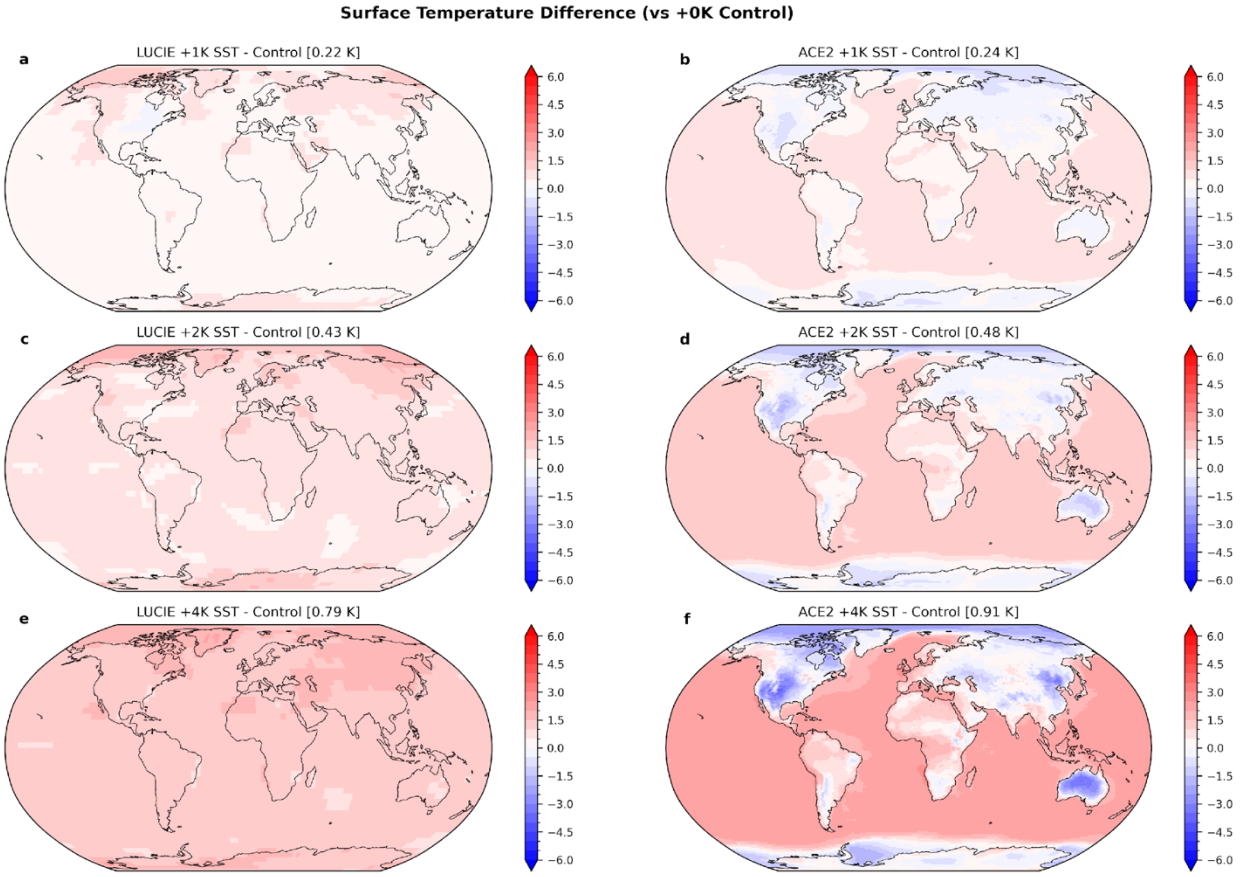

Figure 1. Responses of LUCIE-3D and ACE2 for different SST perturbations. ACE2 shows incorrect sign in the responses while LUCIE-3D does not.

3. LUCIE-3D is trained with a spectral regularizer, which they say mitigates spectral bias and point to past literature. In order to validate this claim, I think the authors need to show that the spectra are the same in LUCIE-3D and ERA5. (I think past literature explores this, but it uses different models, not LUCIE-3D itself.) Based on the PDFs in Fig 8, do the moisture and precip variables still have a spectral bias?

Author's response: We thank the reviewer for this comment. Precipitation and specific humidity indeed still have spectral bias, especially in the high wavenumber range. In the LUCIE-2D paper (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2025MS005152), we have performed ablations to show the impact that the spectral regularizer applied to only the extra-tropics has on both the global spectra and long-term temporal bias. The regularizer improves the spectrum but cannot completely erase spectral bias. A complete erasure of spectral bias cannot be guaranteed

in any system with power-law spectra and definitely not with systems having multiple variables with different power-law spectrums. We have shown a mathematical derivation and cause of this in https://arxiv.org/abs/2304.07029. We have now added the surface and near surface variables power spectra plots and stratospheric power spectrum plots into the result section. The rest of the power spectra plots are added into the appendix.

4. Is the precipitation climatology believable, and is the spectra precipitation blurred? Precip is validated in Figure 6 and Figure 9 (bottom right), which shows that the precip tail is heavily underestimated. Given the underestimation in Figure 9, does the climatology of precip have a sufficiently low bias? And are the precipitation results significantly blurred as measured by the spectra?

Author's response: We thank the reviewer for this comment. We have added the RMSE table into the result section. Precipitation indeed has a noticeable error in the high wavenumber range. We believe this is due to the diagnostic nature of the precipitation. Also, the precipitation error values are significantly lower than variables like near surface temperature. We expect to implement a better weighting scheme in the future for the loss value of precipitation as well as high altitude variables to balance the loss values.

5. To me, this emulator appears to be very similar to ACE2. They both use the same model architecture (SFNO), they both input atmospheric CO2 forcing in the same way, they both have 8 vertical levels with a 6hour timestep, and they both inherit the same stratospheric biases. The main difference is that ACE2 has more diagnostic variables (e.g. turbulent and radiative fluxes) and trains on 1 degree data. LUCIE is coarser resolution, on a T30 grid (~3.8 degree data). This is likely the major reason why LUCIE trains faster than ACE2. Would it be fair to characterize LUCIE-3D as a coarse-resolution version of ACE2?

Author's response: We thank the reviewer for this question. LUCIE and ACE2 are indeed similar to each other but with major differences. LUCIE is trained to predict the tendency of the variables (multiplied by Δt) and ACE2 is trained to predict the full fields. Our experiments with LUCIE-2D (see https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2025MS005152) show improvement on the long term stability and climatology with the current setup. ACE2 incorporates physical constraints including conservation laws in the model. LUCIE is purely data driven with no explicit physical constraints in the model. Furthermore, as newly added into the paper, LUCIE is extended with a probabilistic SFNO formulation thereby allowing for ensemble experiments incorporating epistemic uncertainty.

Minor Comments:

1. NeuralGCM displays a property where some initial conditions are stable and others are not. Does LUCIE-3D show signs of instability if initialized with different initial conditions? Or does LUCIE-3D fully solve this problem?

Author's response: LUCIE-3D is stable with all the initial conditions given to our best knowledge. As shown in Figure 10 and Figure 11, LUCIE can start with unrealistic initial conditions and correct the inference back to the climatology. A major motivation behind several of the differences between LUCIE and ACE2 comes from our theoretical understanding of instabilities in data-driven autoregressive emulators and thus we are confident that LUCIE is a stable model.

2. In Figures 12 and 13, from the caption, I don't understand what the ERA5 data corresponds to, since the x axis is year 0,1,2,3. Is the ERA5 data initialized at 1981 (is it a climatology?) Furthermore, it would be helpful if the Figure 12 and Figure 13 captions stated explicitly if they correspond to the LUCIE-3D variant with SSTs or without. Is the result robust across these 2 LUCIE 3-D variants?

Author's response: We thank the reviewer for the suggestion. The performance of LUCIE-3D and LUCIE-3D with SST are similar in the experiments shown by Figure 12 and Figure 13 and the model shown in the figure is LUCIE-3D **without** SST. The ERA5 global average starts from 1981. Since initialization with climatology and zero fields doesn't present a specific starting year, the x axis is labeled as the year of inference. The caption has now been updated to better reflect this.

3. The authors should clarify the approximate resolution in degrees of the T30 grid. This would be helpful for readers.

Author's response: We thank the reviewer for this suggestion. The resolution in degrees is added to the Dataset section.

4. I was confused by the title and think that there should be a clarification in the text. Based on the title, I thought that LUCIE-3D included 3D architectural components, like PanGu's 3D Earth-specific transformer (many of the readers will likely be familiar with PanGu). However, from the Zenodo codebase, I think the authors are adding the vertical fields as additional 2D fields, and the SFNO operates on the 2D fields with spherical harmonic transforms.

Author's response: We thank the reviewer for this question. By 3D, we mean that the 3D atmosphere is simulated across multiple vertical levels. While the SFNO does not operate on the 2D fields separately, there are MLP layers in between which enforces mixing across the variables. However, we agree with the reviewer that we do not have a 3D Earth specific transformer. LUCIE, despite that is stable while Pangu is not even in the absence of SST forcing (see the LUCIE 2D paper, see https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2025MS005152).

5. Missing citation: I think the paper should cite ArchesWeather and ArchesClimate. These are another class of emulators that is designed to be efficient (and accessible to academic labs).

Author's response: We thank the reviewer for this suggestion. We have now added ArchesClimate to our literature review.

6. A surprising result and contribution of this paper is that it shows that models cannot extrapolate out of sample for boundary conditions, but they can extrapolate out of sample for initial conditions (e.g. all zero initial conditions). Do the authors have an explanation for this behavior? It's of course hard to say why because of the overparameterized nature of ML which can be a black box, but if the authors have any reason why this might be the case, I would welcome more discussion on this in the Discussion section.

Author's response: We thank the reviewer for this insightful comment. I think whether LUCIE-3D or any other class of observation/reanalysis-trained emulators can generalize to out of sample boundary conditions remain to be seen, although we agree that currently it seems to struggle on many fronts. For example, we know that LUCIE-3D does not invert the sign of the responses over land and overall gets polar amplification. Similarly, there are many other aspects of the response that might be thermodynamically incorrect. This is a fundamental ML problem where out-of-distribution generalization remains a challenge without embedding some notion of the change in the model itself. For the initial conditions, we believe that he forcing variables are helping the model to correct the inference back to the attractor. Based on our experiments with LUCIE-2D, without the forcing variables, especially solar radiation, the model won't be able to maintain stability. In another experiment, when the model is given wrong or unrealistic forcing variables, the inference accommodates itself to match with the frequency of the unrealistic solar radiation, for example.

Technical correction

1. Typo: Line 113 on page 5 should be “polar amplification” not “olar amplification”

Author’s response: We thank the reviewer for the correction.

2. Typo: Figure 7 caption “Annualr” should be “Annular”

Author’s response: We thank the reviewer for the correction.

3. Sometimes the authors refer to a figure without actually stating the figure number. This is a small thing but it would be helpful to ensure the figures are referred to in the text directly by number. For example, page 8 and 9 should say “Fig. 6” explicitly.

Author’s response: We thank the reviewer for the suggestion. The manuscript has been updated accordingly.

4. Figure 5 caption itself wasn’t sufficiently explanatory. It should clarify that the interpolated SST output refers to the Gaussian convolution to smooth land and sea discontinuities, and it should point to the relevant section in the text.

Author’s response: We thank the reviewer for the suggestion. The manuscript has been updated accordingly.