Manuscript title: LUCIE-3D: A three-dimensional climate emulator for forced responses
Manuscript number: egusphere-2025-4305
Authors: Guan et al.

My overall recommendation is:

- Major revisions

In brief, my recommendation reflects the need for substantial revisions to address the major and minor comments below. The authors are tackling a genuinely difficult and important problem, and I appreciate the ambition of extending LUCIE into a fully three-dimensional climate emulator. However, in its current form, the manuscript does not yet provide a sufficiently careful or systematic assessment of model performance. Quantitative metrics are used somewhat selectively, some substantial claims are not fully supported, and several aspects of the methods and evaluation are underexplained. It is entirely acceptable for the model to have deficiencies, but these need to be documented and discussed transparently, not only the aspects that work well. With a more thorough and balanced analysis, I believe this work could make a valuable contribution to the ML climate emulator literature.

Comments to the authors

1. Summary of the manuscript

LUCIE-3D is a lightweight, fully three-dimensional climate emulator designed to capture the vertical structure of the atmosphere while remaining computationally efficient and stable over long integrations. Building on the earlier LUCIE-2D framework, it uses a Spherical Fourier Neural Operator backbone and is trained on 30 years of ERA5 reanalysis data on eight sigma levels. The model takes atmospheric CO2 as an explicit forcing variable and can optionally include prescribed sea surface temperature, allowing it to emulate aspects of coupled ocean–atmosphere dynamics. The authors show that LUCIE-3D reproduces large-scale climatological means, variability and forced climate-change signals, including surface warming and stratospheric cooling under increasing CO2 concentrations. It also captures key dynamical features such as equatorial Kelvin waves, the Madden–Julian Oscillation, and annular modes, and can be spun up from idealized initial conditions. Overall, the study presents LUCIE-3D as an accessible, data-driven framework for efficient exploration of climate responses to external forcing while highlighting its potential for future extensions toward more fully coupled emulation.

The manuscript:

- Introduces LUCIE-3D, a three-dimensional machine learning climate emulator based on a Spherical Fourier Neural Operator backbone.

- Trains the model on ERA5 data (8 sigma levels, 6-hourly, T30 grid) with CO_2 (and optionally SST) as forcing.
- Evaluates climatology, forced climate change signals, large-scale variability (e.g. MJO, annular modes), extremes, and spin-up from idealized initial conditions.
- Claims particular novelty in computational efficiency and explicit CO_2 forcing, with potential for coupled ocean–atmosphere emulation.

Please find detailed comments below, separated into major and minor points.

2. Major comments

M1. Insufficient detail in the methods section

The methods section is, in my view, too sparse for a modeling paper in GMD. Key elements of the model design and training procedure are only briefly mentioned or deferred to the LUCIE-2D paper. This makes it difficult for readers to fully understand, reproduce, or adapt LUCIE-3D. I recommend substantially expanding the methods to include:

- A self-contained description of the architecture (input/output variables, SFNO configuration, vertical treatment, and integration scheme).
- Details of the training setup (loss functions, normalization, spectral bias correction, optimizer and learning rate schedule, regularization, and training/validation/test splits).
- A clearer explanation of the Euler integration-based constraint and any other stability-promoting design choices.

These additions could partly be placed in an appendix, but the main text should still provide enough detail for the reader to understand the core design and training choices without needing to consult prior work.

M2. Limited assessment of climatology and model accuracy

The current evaluation of the model climatology, particularly in Sect. 4.1, feels too limited to support the conclusion that LUCIE-3D achieves "good accuracy". For instance, showing a single vertical structure plot, where positive and negative biases can partially cancel when averaged, is not sufficient to characterize the climatological performance. Saying there is "little bias" doesn't really prove this point. I recommend:

- Expanding the diagnostics to include spatial maps of mean state and biases (at multiple levels), as well as zonal-mean sections, for key variables.
- Providing quantitative metrics (e.g. RMSE, pattern correlation, variance ratios) for climatology over well-defined regions and levels.
- Clarifying the time period over which climatological statistics are computed and ensuring consistency across figures.

A more systematic assessment would better support the claims about climatological accuracy and help readers understand where the emulator performs well and where it has limitations. In addition, the discussion of stratospheric representation and the role of the QBO should be tempered. The model has only a single vertical level within the typical QBO altitude range, and the QBO is primarily a tropical signal. Under these constraints, it is not realistic to attribute stratospheric deficiencies to the absence or misrepresentation of the QBO. Instead, the manuscript should emphasize the limited vertical extent and coarse stratospheric resolution of LUCIE-3D as the primary factors shaping its stratospheric performance and be more cautious in drawing conclusions about QBO-related behavior.

M3. Limited and selective use of quantitative metrics

Across the manuscript, model performance is often described using qualitative phrases such as "high accuracy" or "low bias" without accompanying quantitative metrics. In contrast, where the model performs particularly well (e.g., the spatial correlations of the SAM and NAM modes of 0.95 and 0.98, respectively), these strong metrics are highlighted, but even there important caveats are not fully discussed (for instance, the fraction of explained variance differs substantially from ERA5 for each of these modes). It is perfectly acceptable for the model to have deficiencies; however, these should be documented transparently and supported by quantitative diagnostics. I encourage the authors to adopt a more systematic use of metrics throughout (e.g., RMSE, variance ratios, correlation coefficients, explained variance) and to discuss both strengths and weaknesses wherever possible.

M4. Interpretation of the SSW case and implied predictability

Section 4.4 shows a single SSW-like event at 25 hPa and states that LUCIE-3D produces "one of these events in 2006 with inference initialized in 1980". As written, this can be read as implying that the model is expected to reproduce the timing of an individual SSW event many months (or even decades) after initialization. Given that SSWs are generally regarded as having limited predictability on subseasonal scales of at most a few weeks (e.g., Cho et al., 2023, and related work on SSW predictability), this raises several questions:

- What exactly is being claimed here in terms of predictability? Is the goal to show qualitative capability to generate SSW-like events under realistic forcing, or to reproduce the timing of specific observed events?
- What aspects of the forcing or model design would make it reasonable for an SSW to occur in both ERA5 and LUCIE-3D in the same winter, given the long lead time from the stated initialization date?
- If the emulator is primarily intended as a climate model (rather than a forecast system), is it appropriate to emphasize the coincidence in calendar year at all, or could this be misread as evidence of overfitting or overly strong imprint of the training data?

I would recommend clarifying the intent of this example and aligning it with current un-

derstanding of SSW predictability. For instance, you could frame it more explicitly as a qualitative demonstration that LUCIE-3D can produce SSW-like events with realistic structure, and complement this with a more statistical evaluation (e.g., frequency, seasonality, and basic characteristics of SSW-like events), rather than focusing on a single coincident case. Though, now that it is brought up, I suggest clarifying that this is not evidence of overfitting.

M5. Physical constraints and position within the ML emulator landscape

The manuscript does not really discuss physical constraints such as mass, water, and energy conservation, nor how well LUCIE-3D respects these quantities in practice. Other emulators (e.g., CAMulator, ACE2) explicitly include fixers or correction steps to enforce or at least improve physical consistency. It would be very helpful if the authors could:

- Clarify whether LUCIE-3D includes any explicit constraints or postprocessing to address conservation of mass, water, and energy, and if not, provide at least a basic diagnostic assessment of how large the associated drifts or imbalances are over long integrations.
- Discuss how these choices affect the intended use cases for LUCIE-3D (for example, short-term sensitivity experiments versus multi-decadal climate response studies).
- Situate LUCIE-3D more clearly within the broader family of ML climate emulators, in particular relative to models that enforce stronger physical constraints such as CAMulator and ACE2. What niche or role do you envision for LUCIE-3D, given its design choices regarding conservation and physical consistency?

A more explicit discussion of physical constraints and where LUCIE-3D sits in the current emulator landscape would make it much easier for readers to understand when and how this model can be reliably used.

M6. Conclusions and framing of LUCIE-3D's role

The concluding paragraphs strike a good balance between highlighting the promise of LUCIE-3D and acknowledging several key limitations (stratospheric fidelity, lack of dynamic ocean coupling, and sensitivity to prescribed SST perturbations). I would, however, encourage a bit more specificity and alignment with the main body of the paper:

- It would be helpful if the conclusion more clearly articulated what LUCIE-3D is currently well suited for (e.g., idealized forced-response experiments, present-day climate sensitivity tests, process studies focusing on large-scale tropospheric structure) versus applications where it is not yet reliable (e.g., teleconnection studies strongly involving the stratosphere, detailed SSW/QBO analyses, fully coupled ocean–atmosphere variability).
- The statement that LUCIE-3D can ingest SST forcing and produce "physically consistent" atmospheric responses feels somewhat strong in light of the issues documented earlier in the paper (e.g., spurious land cooling under SST perturbations, limited vertical coverage in the stratosphere). I suggest softening or qualifying this wording, or explicitly stating in what sense the responses are physically consistent.

- Since the discussion emphasizes the need for hybrid approaches, improved stratospheric representation, and coupled dynamics, the conclusion could briefly connect this to concrete next steps for LUCIE-3D (e.g., adding physical constraints or fixers, targeted stratospheric training, coupling to an ocean emulator), rather than only framing these as generic goals for "data-driven emulators" as a whole.

A slightly sharper and more concrete conclusion along these lines would help readers understand both the genuine progress represented by LUCIE-3D and the realistic limits of what it can currently deliver.

Technical corrections and typos and places to improve

T1. Page 1, line 10: Replace "(Pathak et al.)" with "(e.g., Pathak et al.)". There are now many ML-based NWP systems, and adding "e.g." makes clear that this is an illustrative, not exhaustive, citation.

T2. Page 1, line 20: Consider adding a more recent and/or peer-reviewed reference in addition to the arXiv preprint (Chattopadhyay and Hassanzadeh, 2023), given that this preprint is now a few years old.

T3. Page 1, line 24: remove (developed by us) or make a more scholarly statement.

T4. Page 2, line 27: I am not fully convinced by the argument that training cost is prohibitive. While non-trivial, a few days on four GPUs does not seem excessively demanding in the current context and could be better justified or rephrased.

T5. Page 2, line 38: The sentence "Unlike its predecessor, which was trained on a limited number of sigma-levels, LUCIE-3D is trained on data spanning the full vertical extent of the atmosphere" is potentially misleading. Because the vertical information is interpolated and the model does not extend above 25 hPa, this is not the full vertical extent of the atmosphere. Please clarify the actual vertical coverage and rephrase accordingly.

T6. Page 2, line 47: The manuscript states that the model has potential for coupling to dynamical ocean models. However, in its present configuration the model does not include the full set of variables typically required for coupling to dynamic oceans. Please clarify what is meant by "potential for coupling" in this context and specify which additional variables or interfaces would be needed.

T7. Page 2, line 55: "Vertically interpolated across eight" is ambiguous. Do you mean vertically averaged (as in ACE) or simply interpolated to eight fixed levels? Please clarify the procedure. Also, please give the nominal horizontal resolution in degrees (e.g. T30, approximately $3.75^\circ \times 3.75^\circ$, corresponding to a Gaussian grid of 96 longitude by 48 latitude points) and briefly comment that this is a relatively coarse resolution.

T8. Page 3, section "Dataset": A small table listing all variables used (including units and number of levels per variable) would improve clarity. Please also explicitly cite the ERA5 dataset. It would be helpful to add a brief motivation for the choice of variables.

T9. Page 3, section "Methods": The statement that SFNO is "well-suited" is rather qualitative. Please provide a more scientific justification (e.g. handling of spherical geometry, spectral properties, ability to capture multi-scale dynamics) and, if possible, support this with references.

T10. Page 3, section "Methods": The description of the model architecture is very brief and relies heavily on the previous LUCIE-2D paper. It would be helpful to include at least a concise, self-contained description of the core architecture here (potentially moving further details to an appendix) so that the manuscript is readable without constantly referring back to earlier work.

T11. Page 3, section 3.1: The "Euler integration-based constraint" is not clearly explained. Is this simply predicting tendencies and then updating the state with an explicit Euler step? Please provide a precise description. At present, the reader must rely on the LUCIE-2D description to fully understand the setup.

T12. Page 3, line 89: Please clarify how the value 0.005 was chosen (e.g. tuning, prior work, sensitivity tests).

T13. Page 3: Please specify whether and how the data were scaled or normalized before training, and describe the training/validation/test split (time periods, fraction of data, etc.). As written, this part of the methods section feels incomplete.

T14. Page 4, line 90: Please provide more detail on the "corrected spectral bias term": what is the exact form of this correction, and how is it implemented in the loss or architecture?

T15. Page 5, Figure 1 and related text: The uppermost model level lies near the lower portion of the QBO region, so only a small part of the full QBO structure (approximately 10–70 hPa) is represented, one level (25hpa). Thus, I would not expect the model to capture a realistic QBO vertical structure at all. It would just be a fluctuation at the top of your last line. I think line 105-109 is just not relevant for that conversation, given the fact that you are not near the QBO height. Also, the QBO is tropical feature. This suggests the discussion has not fully accounted for the actual vertical coverage of the model.

T16. line 113: typo "polar amplification".

T17. Section 4.2: I agree with reviewer 1 that the 0.5 degree bias seems high for the surface temperature trend.

T18. Section 4.2, line 115 for all of these findings are you still refering to figure 2? how can you look at polar amplification in a single line plot.

T19. Section 4.2, line 117. Specific humidity change is in the wrong location in LUCIE with the trend being focused right on the equator rather than in the northern hemisphere tropics. Adjust this comment

T20. **Interpretation of biased SST experiments and smoothing procedure (lines 135–147)**

The discussion of the +2 K and +4 K SST bias experiments raises several concerns. Despite being a nice result, the statement that the model is "numerically stable and physically consistent" seems too strong in light of the clearly unphysical cooling response over Northern Hemisphere land. I would recommend softening this wording or being more specific about which aspects of the solution are physically consistent.

Second, the attribution of this land cooling to prescribed SST fields with land values fixed at 270 K and associated land–sea discontinuities is plausible, but currently presented without direct supporting evidence. The smoothing procedure, mixing SST over ocean and coastal land via a Gaussian convolution and normalization by a smoothed ocean mask,also feels rather ad hoc and is not described in enough detail to be reproducible (e.g., kernel width, definition of "coastal land" points, and whether this is applied during training only or also at inference). Moreover, blending SST into coastal land points is not physically straightforward, since land "SST" is not a well-defined quantity, so it would be helpful to emphasize that this is a pragmatic numerical fix rather than a physically based boundary treatment. Again, this is a nice result, but feels more of an ad hoc numerical remedy than a fully physical solution.

Finally, the claim that the smoothing "improves the response" and the subsequent conclusion that this behavior indicates a broader difficulty for emulators to "extrapolate well outside of their training data" could be made more cautious. It would strengthen the argument to provide simple quantitative metrics demonstrating the improvement (e.g., pattern correlation or RMSE of the warming pattern) and to frame the extrapolation limitation more narrowly in the context of these particular uniform SST perturbation experiments, rather than as a general statement about all emulators.

T21. Section 4.3, line 149: When introducing the Wheeler–Kiladis diagram and the MJO as a key diagnostic for variability, please add appropriate references (e.g., the original Wheeler–Kiladis paper and foundational MJO references) to support this discussion.

T22. Line 152: The statement that LUCIE-3D "closely matches ERA5 in spectral power within the MJO band" is qualitative. From Fig. 6 it appears that LUCIE-3D may in fact overestimate power in parts of the MJO band. I suggest either (i) providing a quantitative metric (e.g., power ratio, integrated variance in the MJO box, correlation across the spectrum) to substantiate "closely matches", or (ii) softening the wording to acknowledge any apparent overestimation.

T23. Lines 153–155: The text states that incorporating the full vertical structure in LUCIE-3D

"grants the model the ability to represent the spectrum of Kelvin waves", in contrast to the earlier 2D version. It would be helpful to clarify what specific deficiency existed in LUCIE-2D (e.g., weaker amplitude, incorrect phase speed, missing parts of the Kelvin band) and to show how LUCIE-3D improves on this with quantitative metrics. Since you highlight MJO and equatorial wave representation in the abstract, this section would benefit from a more systematic analysis, including a brief discussion of remaining deficiencies as well as successes, supported by measurable diagnostics (e.g., power spectra in Kelvin and ER bands, comparison to theoretical dispersion curves).

T24. Line 175; what is the improvement or strategy in Kent et al. (2025)? briefly summarizing which elements of Kent et al. (2025) you have in mind and how they could be applied to LUCIE-3D (for example, a systematic evaluation of SSW frequency, timing, and composite structure, or an analysis of low-frequency precursors), or removing the reference to Kent et al. (2025) at this point and instead making a more generic statement about the need for a dedicated, quantitative SSW evaluation in future work.

T25. Section 4.4 would we expect to have the ability to predict a SSW at 6 Months lead time? If not this is evidence of the model being too fit to the data. What is driving the SSW such that it should show up in both ERA5 and LUCIE at the same time (1980) at 6 months lead? This paper does not seem to indicate that they are at all that predictable. https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023JD039559

T26. Line 202-203. Expand on this sentence or remove.