The authors make both conceptual and methodological contributions to the emulator literature. Conceptually, they frame the climate system as a stochastic process whose evolution can be described either through probability densities or statistical moments, using operator theory and linear response. A key contribution is their demonstration of how different emulation techniques connect to operator and response frameworks, highlighting a current gap: the underexplored role of operator-based emulators. Methodologically, they implement six emulation techniques and evaluate their skill on toy climate models designed to test memory effects, hidden variables, internal variability, and nonlinearities.

Overall, I find the contribution meaningful. Emulator development has so far been largely driven by immediate data needs, and the field lacks a unifying conceptual baseline. This paper takes an important step toward such a baseline and clarifies how approaches from other disciplines fit into the emulator literature. I also find the discussion of pattern scaling valuable, as it helps delineate the conditions under which this widely used technique succeeds or fails.

I see some room for improvement, particularly regarding structure and clarity. Below, I organize my comments along three dimensions: the theoretical framework, the connection between emulators and theory, and the experimental design. Each section starts with a summary of what I think the paper said followed by specific comments. Feel free to use (or ignore) everything as you see fit.

Theoretical groundwork

The paper frames the climate system as a stochastic process whose evolution over time can be described using stochastic differential equations (SDEs). This description allows the dynamics to be restated in terms of operators, which can act either on the probability density (Perron–Frobenius/Fokker–Planck operator) or on observables (Koopman operator). Observables include, but are not limited to, the moments of the distribution (mean, variance, etc.). The two operators are dual: the Fokker–Planck view evolves the density and integrates against the observable to obtain expectation values, while the Koopman view evolves observables directly, with expectations computed with respect to the initial distribution. The operator view allows the problem to be restated as an eigenproblem (e.g., Fokker–Planck view = eigenfunctions are modes of probability densities and eigenvalues represent decay rates). And linear response theory is essentially about considering perturbations to they system and expressing them in terms of solutions to the unperturbed system. More general than the fluctuation-dissipation theorem (FDT) presented in the paper, would be Ruelle's response theory (not mentioned) as it applies to chaotic, non-equilibrium systems. I am not an expert on SDEs, but found the concepts familiar from quantum mechanics (Schrödinger picture vs. Heisenberg picture, Hamiltonian generators, perturbation theory).

Comments / Suggestions:

1. **Framing & readability**: I found the theoretical framing very helpful and intuitive, especially the explanations around L.145 ff. To make it more accessible, I suggest separating the conceptual groundwork from the emulation idea:
    - Introduce a dedicated "Theoretical Foundation" section to present the stochastic framework and operator duality independently of the emulation

target. I think this would help establishing a common baseline first and to invite readers to only read the part they are interested in.
   - Then have a section “Connecting Emulators to Theory” where you introduce what an emulator is meant to do (current 2.2).
   - Finally, present “Experiments/Applications” (current 2.3 + 3).
2. **SDE context**: As I understood, Eq. 1 is a special case derived by Hasselmann under assumptions such as time-scale separation between fast (weather) and slow (climate) variables, stationarity of the fast system, and viewing fast processes as random forcing. A brief explanation of these conditions and the meaning of each term (e.g., the white noise term representing aggregated fast-variable effects) would aid clarity.
3. **Equation consistency**: In Fig. 2 you write $\frac{\partial w}{\partial t} = \mathcal{N}(w, F) + \epsilon\xi$ while Eq. 2 is $\frac{\partial w}{\partial t} = \mathcal{N}(w) + F(t) + \epsilon\xi$. These are not equivalent if feedback parameters depend on forcing. I don’t quite understand which formulation is used when and why?
4. **FDT introduction**: Before directly introducing linear response theory via the FDT (L.219 ff.), an intermediate step would help:
   - Define an unperturbed generator K_0 and perturbation, δK, so $K = K_0 + \delta K$
   - Then the perturbed expectation value of an observable g is $\langle g \rangle_t^{perturbed} = \langle g \rangle_t^{(0)} + \delta \langle g \rangle_t + \mathcal{O}(\delta^2)$.
   - From there, Ruelle’s response theory provides the general solution, with FDT as the special equilibrium case (Eq. 12)

Connecting emulators to the theoretical groundwork

Earth system models generate data that implicitly obeys Eq. 1 (or its operator-based equivalents), but they do not provide us with the exact operators or their solutions. Instead, we observe samples, and the emulator’s task is to approximate either the full distribution or selected observables from these samples. The key conceptual contribution of the paper is to connect emulator-based approaches to operator and response formulations. Modern probabilistic models such as Bayesian inference or diffusion models are naturally connected to the state-based view (Fokker–Planck). For example, diffusion models learn a score function (the gradient of the log density), which directly appears in the Fokker–Planck operator. From this perspective, such emulators can be seen as approximating the Fokker–Planck operator, and thus as providing an entry point to linear response theory: score-based response functions are essentially obtained by applying linear response but replacing the Fokker–Planck operator with the learned score. Similarly, emulators that target specific observables (e.g., the mean of the distribution) connect to the Koopman operator and can therefore be framed within linear response theory as well. The authors explicitly work out these connections for six emulation techniques, focusing on temperature anomalies, and discuss the errors each emulator makes in approximating the underlying operators.

Comments / Suggestions:

1. **Clarity and structure:**
   - This conceptual bridge between emulators and operator theory is central to your paper. I recommend making it an independent section that explicitly highlights

these links and their implications for emulator errors (see comment on Framing in the previous section)
  - An additional table or expanded version of Table 1 would be very helpful. It could summarize pros/cons of each approach, e.g. the computational speed of pattern scaling vs. its structural biases, or the expressiveness of score-based emulators vs. challenges in accessibility and training. In addition, the conncetion of Table 1 to Fig. 2 could be strengthened by adding a column called Emulator type (or adding brackets) that show Method I belongs to Pattern Scaling + Extensions; Method II + IV are impulse response emulators and Method V + VI to Operator-Based emulation
  - Figure 1 and Appendix A1 are excellent in motivating the error sources and in giving an example of how your framework helps identifying them .
2. **Data expectations**: When first reading Section 2.2, I expected the framework to be applied to CMIP6 data (reinforced by Fig. 1). It only became clear later.
3. **Assumptions in pattern scaling vs. impulse response vs. operator-based modelling:**
   - In L. 268 & 299 ff: you righty point out that pattern scaling assumes time-invariance and quasi equilibrium and then mention equilibrium conditions in L. 314 ff. again in terms of FDT. I found the mentioning of two types of equlibrium conditions a bit fuzzy upon first reading and I think it wuld make sense to be a bit more explicit
   - In L. 440 ff. you introduce operator-bsed emualtors as the most general class. This makes sense because the previously introduced emulators have some equilibrium assumptions. I feel like the generalisability of this emulator-based framework could be highlighted a bit more; consider making the assumptions in Fig. 2 (arrow from 3b to 3d) more explicit
4. **Conclusions:** Your reflections in L. 838 ff. are supper fitting. For me, the theoretical contributions were the most compelling part of the paper, since many existing studies implement emulators more naively. As you argue, the lack of a conceptual baseline makes it hard to integrate insights across disciplines, and I would encourage you to highlight this contribution more strongly throughout the manuscript.

Experimental approach (sources of error)

The authors employ four simple climate models (a two-box model, a three-box model, a noisy box model, and a cubic Lorenz system) and drive each with four structurally distinct forcing pathways (abrupt, transient high-emissions, plateau, and overshoot). For each experiment, they train their emulators on data from one scenario and test them against all other scenarios, thereby comparing performance across settings. The experimental design is deliberately simple: it targets errors arising from dynamical features such as memory, hidden variables, noise, and nonlinearities, rather than errors linked to spatial heterogeneity. This choice has clear advantages—the box-model experiments are transparent, reproducible, and well-suited for stress-testing emulator failure modes—but it also carries disadvantages, as spatial patterns are not represented and emulator performance is reduced to a single aggregated metric across boxes.

Comments / Suggestions:

1. **Data expectations**: Reiterating the previous point. Initially, I found the experimental set-up somewhat confusing, as I expected the tests to involve spatially resolved ESM data. The title, abstract, and introduction explicitly mention ESMs; Fig. 1 also presents spatially resolved data; and in Section 2.2 the term “spatial” led me to expect an evaluation of spatial error. In practice, however, the experiments are based on box models with at most three degrees of “spatial” resolution. Applying your framework to toy models is, in my view, valuable—as it provides a controlled setting for isolating and examining specific effects—but I think this would be clearer if explicitly framed as a proof-of-concept.
2. **Averaging of results**: The data only ever shows a single evaluation score, while sometimes the models have multiple boxes. Do you average across boxes?
3. Table 6 summarizes the experimental findings well.
4. I appreciated the conclusions in L. 816 ff.

Other minor suggestions

- Fig. 2:
  - Add a reference to Table 1 to the description (helpful for understanding the boxes on the right given you refer to Fig.2 already in L.77, but mention the Methods I-VI only from L.115 onwards)
  - What is the difference between solid and dashed arrows (e.g., going from 2b to 3b as opposed to 3b to 3d)?
- L. 759: Non-linearities as opposed to nonlinearities throughout the remainder of the manuscript