This manuscript investigates the relationship between global surface air temperature (GSAT) anomalies and effective radiative forcing (ERF). The author estimates a power-law with $GSAT \propto ERF^{1.9}$ and projects mid-century warming under high ERF that exceeds the IPCC AR6 best estimate by ~0.6°C.

The author is to be commended for several aspects of this submission:

- Exemplary open science: fully documented Python code & data on Zenodo make the analysis completely reproducible.
- Econometric diligence: The paper makes a serious effort to address challenges such as unit roots, cointegration bounds testing, and attenuation bias. These are too-frequently overlooked in our field.

However, the conclusion that modern warming reflects $GSAT \propto ERF^{1.9}$ and that such a relationship will extend to the future and result in substantially higher mid-century warming, is not supportable in my view.

The power-law formulation is not translation-invariant and depends entirely on an arbitrary forcing baseline. The claimed equivalence to a two-layer EBM breaks down physically, relying on a 3-year mixed-layer relaxation to extrapolate multi-decadal warming where deep-ocean heat uptake occurs. The expanding-window out-of-sample validation does not actually test the power law performance. Conditioning the regression on ~1977 involves uncorrected multi-testing and physically unjustified selection criteria.

I do not believe that the central results can be salvaged by recalibration with the data used here, so my recommendation is rejection. I provide my major concerns below, and follow with some minor comments. The final part of the major comments recommends a route forward, but it is not certain that a paper that addresses my concerns would be achievable in typical review turnaround timeframes, and the methods and conclusion may be unrecognisable.

Summary of major concerns:
The headline result I will label R1: "recently GSAT is best described as a power law of ERF, and the result is potentially ~0.6 °C additional warming by ~2050, compared with the AR6 best estimate".

R1 is shown in Fig. 6 and 8, where the model predicts GSAT around 2050, given an ERF of +0.6—2.1 W m^{-2} over the present day.

To accept R1, I should require that: (i) the model is at least as convincing as alternatives, (ii) that the values and their uncertainties are appropriately estimated, and (iii) that the model can be used to project GSAT response to ERF values that are, say +0.6—2.1 W m^{-2} outside the fit data.

The three sections below will try to explain why I believe there are issues with all three requirements.

1. **Convincing-ness of the model**

Unlike polynomial fits with integer exponents, the power-law model is not translation invariant with regards to ERF, so predicted GSAT depends on the chosen ERF baseline period or year.

The model is effectively constructed so that GSAT responds within ~3 years of an ERF change. A model with such a rapid response is unlikely to have "memory" of what happened hundreds of years ago, so I don't see the justification for using 1750 as the baseline for ERF. If I change the baseline to 1977 to represent conditions during the regression period, ERF shifts by 0.88 W m^{-2}. There are now more years with negative forcing and undefined GSAT, and the Fig. 8 high-GSAT estimate shifts from ~3.0 °C to ~2.6 °C. Both the change in which parts of the sample are defined, and the shift in predicted GSAT in response to an arbitrary definition choice are concerning.

The physics are also hard to interpret. Given:

$$GSAT(ERF_t) = \frac{\kappa T_D}{\lambda + \kappa} + \frac{F_{ref}}{\lambda + \kappa}\left(\frac{ERF_t}{F_{ref}}\right)^{\nu}$$

If $ERF_t \to 0$ is equilibrium, then the derived intercept is:

$$GSAT_{eq} = \frac{\kappa T_D}{\lambda + \kappa}$$

With $\lambda > 0$, $GSAT_{eq} < T_D$ and the model implies an upward (cooling) ocean heat flux at equilibrium, which depends once again on the definition of the ERF baseline. The argument is that this simplified model captures some transient feature of GSAT(ERF), but also that this "short term" model can be fit over ~50 years of data, and predict decades more into the future. This is deeply inconsistent.

The model lacks physical basis for choices, it cannot capture negative ERF, its final results depend on baseline choices which appear arbitrary, & it is explicitly simplified for short-term modelling, then used to project out over almost a century.

2. **Model parameters and uncertainties, and prediction uncertainties, do not appear to be appropriately estimated**

The paper does an admirable job of identifying and testing potential statistical pitfalls, but I don't think it actually quantifies the uncertainties of its final results. I believe there are *two* uncertainties here:

i. What are the uncertainties in the present-day model fit (in parameters and its predictions within the sample)?
ii. What are the uncertainties in projecting to ERF outside of the training sample?

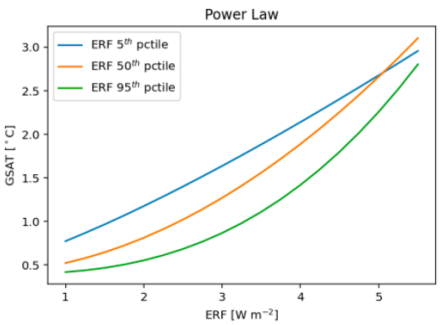
Idiomatically I feel that much of the current manuscript misses the forest for the trees; some elements of (i) are intensively studied, while other components are missed. Crucially, I don't see how (ii) is tested at all.

The year-at-a-time refit/prediction step covered on L288—308, L491—511, L638—650, and Figure 5 is 2+ pages of content that doesn't clearly support R1. The power law parameter seems to be the one fit over the full period, the evaluated year-to-year performance stats are likely dominated by the GSAT persistence + ONI terms. The ability to predict year-to-year changes within the window doesn't tell us much about the power law parameter uncertainty, and it doesn't address (ii) at all. The jackknife tests do seem to evaluate ν response, but one-at-a-time removal assumes independence in the data when we know there's strong autocorrelation and cross-correlation.

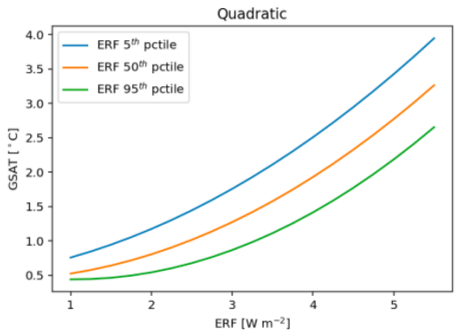
Regarding (i): I think there are other factors.

- Multiple start dates are tried; this is effectively multiple trials and I see no correction for having tried and selecting the "best" – selecting a start point because it gives the best statistics results in overconfidence. Furthermore, I do not follow the argument for even selecting the narrow window within which the start is allowed.
- The GSAT and ERF data are both autocorrelated, but I can't see how the ν uncertainties (and subsequent GSAT projections) account for this.
- The SIMEX correction assumes that ERF uncertainty is also independent, although L607—611 correctly notes that ERF errors are likely to be persistent.

I am very happy to see efforts made to cover attenuation bias in regression: our community benefits from this. However, I believe the SIMEX correction assumed that ERF was iid. Below I checked the ERF = 5.5 W m^{-2} GSAT prediction using the 5th, 50th and 95th percentile ERF time series, which is a more realistic representation of their uncertainty (as noted on L607—611). More-realistic errors in ERF result in reduced projected GSAT at the highest ERFs considered – the opposite of what one would expect from (nonphysical) independent random noise.



Along with the other problems I discussed with the power law, the power law fit says that uncertainty in ERF leads to a tightening of uncertainty in GSAT for sufficiently large values of ERF – values which are far outside the fit sample and are very far away from the "transient" behaviour that the model is derived for. This is yet more evidence for the inappropriateness of the power-law fit being used for longer-term projection. By contrast, a quadratic fit shows uncertainty broadening as ERF increases, which makes far more sense.



Related to the multiple-trials issue and the selection of the regression period: I do not see why 1977 should be selected as a start. Arguments related to a Pacific phase shift and "coherent" forcing are presented, but the following choice of year appears to be a non-sequitur to me.

There are numerous identified post-1977 North Pacific shifts (e.g. Overland et al doi: https://doi.org/10.1016/j.pocean.2008.03.016). Is the power law regression sensitive to N Pacific phase? If yes, then are we only projecting forward assuming no similar shifts until 2050? And what makes 1977 sufficiently different that the other ones identified by Overland or similar ones visible in time series of PDO index during 1977—2025?

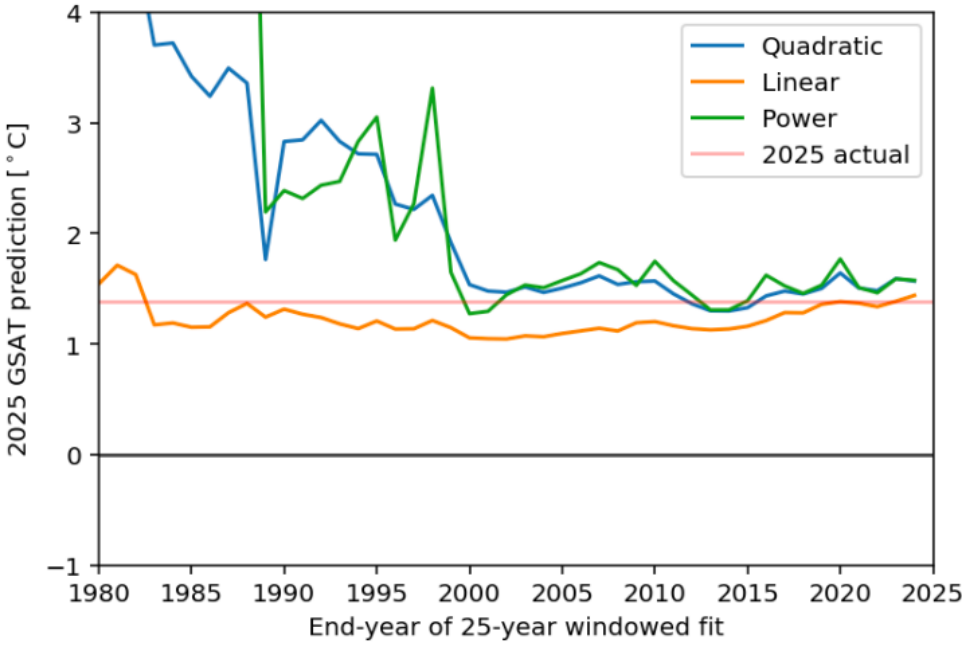
The argument that we should begin in 1977 also mentions that forcing should be "coherent". I don't see why that is a requirement – are we saying that the model only works when forcing agents are all moving in the same direction? There are changes in the relative strengths of GHG and aerosol through time too, and the ratio of GHG:aerosol forcing changes in future scenarios too.

In summary; while a lot of statistical efforts were made, I don't think they appropriately accounted for the effect of multiple trials in selecting the data period, or non-independence of the data in terms of quantifying uncertainty. I really commend the attempt to account for attenuation bias, but its assumption of independence reveals nonsensical behaviour of the power-law fit, and might even result in a correction of the wrong sign. The physical justifications for the selection of the period also don't convince me.

3. **Does a power law GSAT(ERF) constrained in one period provide information about longer-term responses to large changes in ERF?**

This is a fundamental question that I do not see addressed. It is argued that the model is measuring some transient property of the system, and that these transient properties can predict long-term changes.

I tried fitting 25-year windows and predicting 2025 GSAT below. There are worryingly wild errors prior to fits ending around 2000, and even then the quadratic model seems similarly good at predicting future GSAT – and the major conclusion of this manuscript is related to predicting future, out-of-sample GSAT. No number of tests performed on in-sample data can truly provide this sort of test or justification.



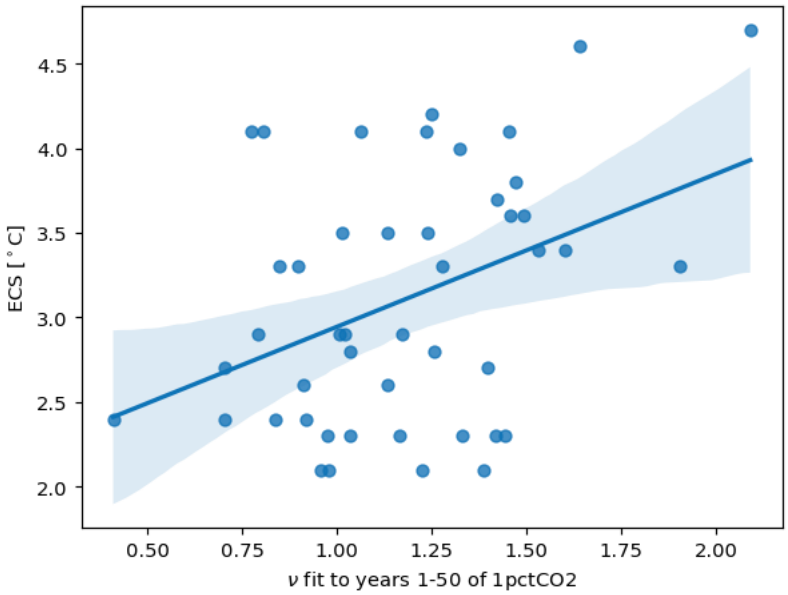
4. Summary/possible future approaches

The paper is quite complex and challenging to follow – so I may have missed some details. But to be convinced I would like to see (i) reliable estimation of the model uncertainties within the fit period, and (ii) evidence that GSAT(ERF) during some selected period is informative about future warming.

I'm thinking of approaches somewhat like emergent constraints (although this is a minefield, e.g. Caldwell at al., 2018, doi: 10.1175/JCLI-D-17-0631.1) mixed with a more-rigorous and consistent treatment of uncertainties informed by the climate's behaviour (e.g. large ensemble Monte Carlo; Glaser et al, 2024, doi: 10.1088/2752-5295/ad69b6).

As a simple test, I looked at 1pctCO2 simulations from CMIP5 and looked at whether a power-law fit to the first 50 years is predictive of their equilibrium climate sensitivity. I just did this as a method demonstration: the ERF is nearly linear in time, so the regression is trivial. I accept that it's a poor representation of reality.

One can squint and decide whether there's a likely relationship there, but it doesn't look hugely informative to me – it seems unlikely that knowledge of a power-law fit will meaningfully tighten uncertainties around future warming within these models. In order to believe that a ~50 year GSAT(ERF) regression is informative about long-term warming, I'd want to see some evidence analogous to the below. It would additionally require more-realistic scenarios and/or physics, that justifies the use of short-term fits to project long-term warming.



MINOR COMMENTS:

- For this sort of application, I would recommend the newer Relative ONI (RONI) rather than ONI. The ONI moving baseline results in a bias at the end of the series whenever there's a trend.
- While I think I managed to work out what was done, I found it hard to digest and cross-reference steps. This paper would probably have benefited from supplementary tables that separately listed the models used for major conclusions (e.g. the power law is fit first, then within-sample annual predictions use GSAT_t-1 and ONI, then the long-term warming sets ONI to zero?) followed by a supplementary table listing the tests done on each one, columns might be something like: test – purpose – assumptions - result. That would have eased my cross-referencing.