ESD-2020-48 Minor Revision

Reviewer 1:

- This paper addresses the same problems as Hebert et al. (2021), but uses the FEBE model instead of a truncated power law. I would like to see some more comments on the strengths and weaknesses between these two models, and some discussion on what this paper adds to Hébert et al. (2021) to justify this work as a standalone paper.

As mentioned in the paper, Hébert et al. (2021) tamed the divergences by cutting off the power law CRFs at small scales. The caveat was that the CRF model truncation was somewhat ad hoc, and therefore only useful at decadal or longer scales, while the FEBE and its Green's function covers all ranges of scales. While the low frequency Green's function can be very close to Hébert et al. (2021)'s truncated power law CRF, the high frequency regime is able to produce internal variability coherent with the observed scaling and fractional Gaussian noise used for skillfully forecasting the stochastic (internal) variability at monthly, seasonal, interannual (macroweather) scales (Lovejoy et al. 2015; Del Rio Amador and Lovejoy 2019, 2020). In addition, it is much more sensitive to the volcanic forcing and the parameters are more strongly constrained. A significant consequence and improvement over Hébert et al. (2021) is the error model was not ad hoc, rather predicted by the model itself: the internal variability response to white noise internal forcing. The differences in the two models are more thoroughly covered in line numbers 65-70, 87-92, 365-375, 691-700 of the first revision.

Lovejoy, S., Del Rio Amador, L., and Hebert, R.: The ScaLIng Macroweather Model (SLIMM): using scaling to forecast global-scale macroweather from months to decades, Earth System Dynamics, 6, 637, 2015.

Del Rio Amador, L. and Lovejoy, S.: Predicting the global temperature with the Stochastic Seasonal to Interannual Prediction System (Stoc-SIPS), Climate Dynamics, 53, 4373–4411, https://doi.org/10.1007/s00382-019-04791-4, https://doi.org/10.1007/s00382-019-04791-4, 2019.

Del Rio Amador, L. and Lovejoy, S.: Using scaling for seasonal global surface temperature forecasts: StocSIPS, Climate Dynamics, 2020.

- In line 298 the paper argues that the fractional Gaussian noise approximation of the residuals allows the model to take into account the strong power law correlations, but it's accuracy is weaker on low frequencies which only weakly influence the likelihood function. I would like some clarification on what motivates the FEBE (on the applications considered in this study) instead of simply using a power law.

Described in the comment above, the FEBE is a model that can be derived rather than being ad hoc as the case would be when using a pure power law. To make more realistic models, the key issue is energy storage. Storage is a consequence of imbalances in incoming short wave and

outgoing long wave radiation and it must be accounted for in applications of the energy balance principle (Trenberth et al., 2009). As pointed out in Lovejoy (2019, 2021a) and developed in Lovejoy et al. (2021) it is sufficient that the scaling principle not be applied to the Greens (Climate Response) Function, but rather to the storage term in the EBE. In lieu of the energy being stored by uniformly heating a box, energy is instead stored in a hierarchy of structures from small to large, each with time constants that are power laws of their sizes. This conceptual shift can be implemented simply by changing the integer order of the storage (derivative) term in the EBE to a fractional value: the Fractional Energy Balance Equation (FEBE). While Lovejoy et al. (2021) derived the FEBE in a phenomenological manner, Lovejoy (2021b, 2021c) showed how it could instead be derived from the continuum mechanics heat equation used in the Budyko-Sellers models. Indeed, by extending Budyko-Sellers models from 2D to 3D (i.e. to include the vertical) and imposing the (correct) conductive – radiative surface boundary conditions, one immediately obtains fractional order equations for the surface temperature. In other words, nonclassical fractional equations and long memories turn out to be necessary consequences of the standard Budyko-Sellers approach.

Trenberth, K. E., Fasullo, J. T., & Kiehl, J. (2009). Earth's global energy budget. Bulletin of the American Meteorological Society, 90 (3), 311–324. https://doi.org/10.1175/2008BAMS2634.1

Lovejoy, S. (2019). Weather, macroweather and climate: Our random yet predictable atmosphere. Oxford U. Press.

Lovejoy, S. (2021a). Fractional relaxation noises, motions and the fractional energy balance equation. Nonlinear Processes in Geophysics, 2021. https://doi.org/10.5194/npg-2019-39

Lovejoy, S., Procyk, R., Hébert, R., & Del Rio Amador, L. (2021). The fractional energy balance equation. Quarterly Journal of the Royal Meteorological Society, n/a(n/a). https://doi.org/https://doi.org/10.1002/qj.4005

Lovejoy, S. (2021b). The half-order energy balance equation, part 1: The homogeneous hebe and long memories. Earth System Dynamics Discussions, 1–36. https://doi.org/10.5194/esd-2020-12

Lovejoy, S. (2021c). The half-order energy balance equation, part 2:the inhomogeneous hebe and 2d energy balance models. Earth System Dynamics Discussions, 1–44. https://doi.org/10.5194/esd-2020-13

- There is currently very little description on the limitations of the FEBE. The manuscript would benefit from further discussion on where the model is suitable and where it is not.

The FEBE is in fact a regional (horizontal space) – time model that here is integrated over space to yield a "zero-dimensional" model similar to the standard "Box model" except for a different order of differentiation. At the moment it is linear, but nonlinearities such as temperature-albedo feedbacks are easy to introduce. Other temperature- forcing feedbacks such as those responsible for tipping point phenomena could also be easily incorporated. Since the FEBE can be generalized in many ways, its limitations are not in fact clear. A practical difficulty (but not a limitation) is that some of the parameters- especially in the regional FEBE are difficult to empirically estimate (especially the regional relaxation times). These issues will be explored in further publications.

- In the estimation of the model parameters the forced response is subtracted from the data before the residuals are fitted using Mathematica. In my understanding this is done by first sampling parameters from the prior distribution, then by removing the corresponding forced response (based on the simulated parameters) before fitting a fractional Gaussian noise process to the data. However, since the removed forced response also depends on the same parameters which are to be estimated, this component is not fitted to the data before it is removed and hence its shape is determined entirely by the priors. This could make the model more sensitive to the choice of priors.

The parameters are not initially sampled from the prior distributions, but rather sampled from broad uninformative uniform distributions to generate a wide array of possible forced responses. Then we obtain many residual temperature series by removing the generated forced responses from the global temperature series which is a fraction Gaussian noise process as predicted by the FEBE model itself. From these residual series, we calculate the likelihood of being a fractional Gaussian noise with parameter h, giving us our multidimensional likelihood function wholly independent to the prior distributions which are applied later using Bayes.

In my opinion, it would be better if both the forced response and the residuals were to be fitted simultaneously. This can be achieved by e.g. a hierarchical Bayesian modeling approach. This framework is also able to incorporate non-Gaussian priors, which could possibly remove the need for approximating the joint posterior.

We understand what the referee is suggesting, but this is an unnecessary complication as the error model for each possible forced response is given theoretically by the FEBE itself (when the FEBE is forced by a white noise, as is the case in Eq. 1, the stochastic portion is a fractional Gaussian noise – the residual series). The joint posterior is only approximated by a Gaussian for computationally efficiency when generating many realizations of projections.

- I would like to see a comment added to the text which ensures that the Gaussian approximation of the joint posterior (Eq. (21)) is indeed accurate.

The Gaussian approximation was only used for projections, rather than creating a net over the large five-dimensional parameter space to draw parameters (as was done in Hebert et al. 2021 – a computationally expensive process). The parameter distributions for all five parameters shown in the results (Section 3) are all from the marginal probabilities of the joint posterior – their Gaussian appearance and computationally considerations are why a Gaussian approximation was chosen.

- Gaussian priors imply a non-zero probability of negative values, which could cause e.g. scaling parameters to be negative. I would like a comment that addresses if/how the authors have constrained the model parameters.

For the model parameters, h and τ, we restrict them to being greater than zero – in the case of the scaling exponent this is justified as the theory of fractional Relaxation noises (fRn, the generalization of fGn) is only for h>0, and a negative relaxation time has no physical meaning (breaks causality).

- In line 467 the authors state that they have performed 500 Monte Carlo simulations of the projections. Has it been verified that the accuracy is sufficient? A comment clarifying this would be welcome. Furthermore, would it be computationally feasible to increase this number, if needed?

The 500 Monte Carlo simulations chosen is already far more than needed. Included is a table showing the mean and standard deviation of the RCP parameters depending on the amount of simulations drawn from the Monte Carlo simulation of the approximated posterior distribution – it is clear that at already 100 simulations we can reproduce the values presented in Table 1 (shown in the first row of this table) in the paper.

Parameters MCMC Simulations	h	τ	α	ν	s
	0.38 (0.05)	4.7 (2.3)	0.60 (0.40)	0.28 (0.13)	0.56 (0.11)
10	0.41 (0.05)	5.1 (2.0)	0.44 (0.27)	0.32 (0.10)	0.48 (0.09)
100	0.38 (0.04)	4.5 (1.9)	0.58 (0.36)	0.31 (0.13)	0.55 (0.13)
250	0.38 (0.05)	4.6 (2.0)	0.62 (0.39)	0.31 (0.12)	0.56 (0.10)

500	0.38 (0.05)	4.6 (2.4)	0.63 (0.41)	0.29 (0.12)	0.55 (0.11)
1000	0.38 (0.04)	4.6 (2.2)	0.64 (0.42)	0.28 (0.11)	0.56 (0.10)

Other than this I have some minor/technical comments and suggestions:

- In the Bayesian framework one should use "credible intervals" instead of "confidence intervals".

- Appropriate punctuation after equations:

(3), (4), (11), (20), (22)

- Figure 13: Caption states that CMIP5/6 MME is represented by black, but in the figure the color is gray.

- Line 693: "Latter" is used when the preceding sentence only has one object

- Line 712: "Projections through to 2100"

- Line 726: "The FEBE could be also" to "The FEBE could also be"

- Line 731: citation should be parenthetical

These will be corrected in the revision.

Reviewer 2:

1. I would suggest to take care of some typos, to close some parentheses, correct figure captions when describing lines or symbols, and to fix some issue as capital letters without any punctuation before. This especially occurs after equations.

Ok.

2. Line 238: "We consider the standard assumption about internal variability that it is forced by a Gaussian "delta correlated" white noise". Could the authors add a reference to this? Why not to use a red noise spectrum that is also generally related to the internal noise?

The internal variability produced by FEBE is forced by a white noise, but does not result in a white noise (the result is red). The internal variability generated is a fractional Relaxation noise, which is a type of red noise. This is the same concept as in Hasselman (1976) where a white noise was used to produce a red noise, but with an exponential function rather than the FEBE. The physical concept is the same: the white noise corresponds to the high-frequency atmospheric

forcing, and the response function (FEBE in our case, exponential in Hasselman 1976) corresponds mainly to the mixed-layer of the ocean which integrates those fast variations to produce a red noise spectrum.

Hasselmann, K., 1976: Stochastic climate models. Part I. Theory. *Tellus*, **28**, 473–485, https://doi.org/10.3402/tellusa.v28i6.11316.

3. Figure 10: I would like to suggest the authors to comment more on this interesting result. In particular:
what is the range of scales over which the scaling exponent is evaluated?
if, as expected, in the macroweather regime temperature fluctuations decrease, I was wondering why the authors do not use the full range of scales belonging to this regime?
If using the (full) accessible macroweather regime to compute h, how to explain the difference with observations? I would expect h~0 for observations.

We do not use the full range of scales to calculate the scaling exponent because the data is a superposition of natural variability and forced response , the latter is not scaling (and in fact, the former is not perfectly scaling either once we approach the relaxation time (here ≈ 5 years). Both effects break the scaling as indicated. The straight line shown in Figure 10 is not a regression, but a reference line to show the theoretical high frequency result when forcing by a white noise internal forcing using the empirical h≈ 0.4 value; over the macroweather regime the observations are as the referee says h~0.

4. Figure 12: it seems to me that FEBE lowered projection uncertainties but also directly projections. How the authors explain this? Could be due to the parametric uncertainty that differs from structural one of MME?

The key reason for the lower temperature projections as compared to the CMIP MMEs, is the lower ECS of the FEBE which is a result of the long memory storage and inclusion of the aerosol scaling factor that accounts for the overly strong historical aerosol forcings. Due to these two main factors, the FEBE has a lower median (and more constrained) estimate of ECS in comparison to the CMIP MMEs, thus leading to lower projection uncertainties and lower projected temperatures.

5. Figure 13: how the authors explain the oscillations observed for the RCP 2.6/SSP 1-26 scenario for FEBE? It seems to me that they are 10-yr period, is there any relation with the transition for scaling laws?

The oscillations observed in the RCP 2.6/SSP 1-26 are caused by the 11-year solar cycle, which can be seen in the same scenario projections in Figure 12. These oscillations also occur in the projections using the higher emission scenarios, although due to the scale of temperature change due to anthropogenic warming they are not visible.

6. Figure 13: it seems to me that there is a good agreement with RCP 8.5/SSP 5-85 scenario for FEBE and MME (although shifted, why?), while a different slope of the probability is found (steeper for FEBE than MME). Could the authors argument on this? Is it related to the intrinsic parametric uncertainty?

In the case of RCP 8.5/SSP 5-85 (along with other scenarios), the FEBE probability of crossing thresholds 1.5K/2.0K (if they are to be crossed) later as compared to the CMIP MMEs. This can be understood from Figure 12, which shows that the warming projected by FEBE is less than the CMIP MMEs, and in the cases where emissions continue to rise (all but scenarios RCP 2.6/SSP 1-26) this will result in the crossing of said thresholds in the inevitable future. The slope may be slightly steeper in the FEBE probabilities as the referee notes but only very tangentially as the RCP8.5/SSP 1-26 probabilities (circles – also the probability curve furthest to the present in all cases) are nearly parallel. The marginal difference in the slopes of the probability curves may be that future aerosols never are reduced to zero (see Figure 1 bottom); so that when the FEBE reduces aerosol forcing by the aerosol linear scaling factor they have a much weaker (nearly negligible effect) in the far future in comparison to CMIP models which maintain a constant cooling forcing into the future.