Response to Reviewer #1

We agree with the reviewer that this discussion has digressed very far from the L&V paper and from our comment to it. But that is not our fault. Rather than sticking to the real issue, L&V started the discussion by publishing a very lengthy reply where the main message was that we were ignorant about the multifractal formalism. We had no other choice than responding to this attempt to discredit us as incompetent novices to the field. It is rather ironic that L&V in their last reply complain that we are "too mathematical," whereas they are "physical." Our comment makes use only of elementary mathematics, we illustrate our points with simple demonstrations, and we mostly use the methods of and L&V and software downloaded from Lovejoy's web site.

The reviewer writes: " I found these public exchanges a bit astonishing: recall that the entire debate is about the degree of linearity of outputs of two indisputably nonlinear numerical models."

How can there be *a degree of linearity*? The meaningful question is: *what is the degree of nonlinearity*. And we never disputed that the models are nonlinear. The issue is if this nonlinearity is strong enough to be detectable *in global temperature* by the methods employed by L&V. The proper way to test this is to find data and methods to reject the linearity hypothesis.

The reviewer writes: My impression is that L+V conceded too little – e.g. they essentially ignored the linear oscillator result as being irrelevant – whereas R+R concluded too much – that this result somehow forces us to accept that the model that we know to be nonlinear is in fact linear."

We find it very depressing that the reviewer, after all the exchanges on methodology in the discussion, can write that we conclude that the model is linear. This is wrong! We do not conclude that, neither in the comment, nor in the ensuing discussion. We have stressed again and again that linearity is a statement that cannot be verified (just like the statement that the photon has zero mass). This is not semantics, it is a fundamental principle in the philosophy of science. But linearity can be falsified by proper data and a proper test, and this is why the linearity hypothesis is a well-posed problem. Our whole point is that the two tests devised by L&V do not reject the linearity hypothesis. If nonlinearity cannot be detected by rejection of the linearity hypothesis in global temperature data, it gives credibility to linear modeling of the global temperature response. This is why it is important to clarify whether L&V's tests are correct or not.

In our Figure 2, we demonstrate that L&Vs test for subadditivity in the ZC-model is invalid. As a reviewer, one of us pointed out the unsatisfactory way L&V dealt with this issue in their Figure 3 of their paper, and this was still unsatisfactory in their published paper. The demonstration in Figure 2 of our comment was not presented in the review process, and this comment is our only possibility of publishing it. Neither the discussion, nor the reviewer, have disputed the correctness of our Figure 2. Hence, if ESD will not publish this demonstration the journal sends the false message that the peer-review has established that Figure 3 of the L&V paper is correct, and that our Figure 2 is wrong. Alternatively, it sends the message that the journal will not accept comments on papers published by influential scientists.

The problems connected with the invalid implicit assumptions (I-III) in the L&V paper were raised by us in the reviews, but rejected by L&V with highly unsatisfactory arguments. We should probably have recommended rejection, but there was obviously a need for a more in-depth discussion of these points. As an alternative, the editor suggested to write a peer reviewed comment, which we did. In Sect. 3.2 of our comment we demonstrate theoretically that imperfect scaling in response function and structure functions may give rise to different estimated intermittency in forcing and response, even if the response is linear. The reviewer does not point out any error in this section.

Nevertheless, we stress in our comment that Sect. 3.2 is not essential for our conclusion. The essential thing is the demonstration that a linear response model with internal noise can reproduce the trace-moment results of L&V. The reviewer buys the argument of L&V that this is demonstrated only by one realization of this linear model, and ignores that we have made the code available for anyone to check that this is a statistically robust result. If the editor invites us to submit a revision, we will include the results of an ensemble run which proves this point.

The most important effect that produces the observed change of intermittency between forcing and response is probably the high internal variability. It is quite astonishing that this effect is not commented by the reviewer. The main reason for including the data from the NorESM model was to demonstrate the crucial effect of that internal variability. How can the reviewer ignore that!

The problem that L&V neglect internal variability was raised in the review, but it was rejected by L&V as unimportant by very obscure arguments. We have demonstrated in our comment (and we will include the NorESM results in a revised version) that it *is* important. If the reviewer cannot prove otherwise, and still recommends that it is not worth publishing as a comment, the reviewer in fact recommends ESD to refrain publishing corrections to incorrect published results.

The reviewer ends his report by encouraging all the authors “to spend their efforts clarifying, quantifying the weakly nonlinear parts of the models and indeed of the real world!” That is an issue we are already working on with data from the NorESM model. However, those nonlinearities are difficult to detect in the global temperature series. The likelihood of detection is much greater in regional or local temperature data, and in other climate variables. One cannot quantify nonlinearity if one is not able to detect it. And detection means rejection of the linearity hypothesis.

The reviewer seems to be of the opinion that it is unimportant whether the tests devised by L&V to detect nonlinearity are valid or not – since the models obviously are nonlinear anyway. Then one cannot avoid asking the crucial question; what was the point of publishing the L&V paper in the first place?

Response to Reviewer #2

We agree with all remarks by this reviewer and have incorporated the necessary changes in the revised manuscript. All essential changes have been marked in red in the revision.

Response to Reviewer #3

We thank the reviewer for illuminating comments, which had been incorporated in the revised paper. Below follows a detailed response to the comments.

[1] *On the hypothesis testing for subadditivity.* We perform two tests for the subadditivity of the ZC model. One, described in Sect. 2.2 and 2.3, is a different test from the one performed by L&V, since it includes internal variability (which L&V do not). The reviewer is quite correct in pointing out that this test does not represent a direct rebuttal of the L&V test, if one assumes that internal variability can be ignored. However, in Sect. 2.4 we repeat the L&V test (i.e., without including internal variability) and present the result in our Fig. 2. We don't find the factor 1.5 claimed by L&V on the longest time scales, so this figure presents a rebuttal of the L&V test. See also our response to Point [7].

On the language - personal and subjective. Lines 164-165. We think it is important to point out that the approximation in question is unnecessary. This is neither personal nor subjective. Approximations are justified if they simplify things, and do not introduce biases. In this case, the approximation does not simplify anything (computations are just as easy without it), and it introduces a bias towards subadditivity. In their revised paper L&V present results both with, and without, this approximation (see their Fig. 3), which demonstrates this bias. We take that as an admission that the analysis based on this approximation is wrong (biased). However, they don't draw the obvious conclusion and omit the approximate analysis and results, but present both as two alternative approaches, and in the concluding section they present the difference between the approximate and exact result in a way that misleads the reader to interpret it as an uncertainty range. We think it is appropriate to point out these facts, but in the revision we have reduced this paragraph to pointing out the nature of their approximation, a brief description of the results they have presented, and a description of our findings.

Line 318: In the revision we have removed the the offending phrase, which we agree is unnecessary.

[2] Line 31: In 8 out of 16 models studied by Geoffroy et al. (2013) one can observe a very small overestimation of the transient response in the 1 pctCO2 scenario when parameters in the two-box model are estimated from the 4xCO2 step-function scenario. This discrepancy does not have to arise from nonlinearity, however. It is just as likely a result of the simplicity of the two-box model. It is well known that a long-memory response will lead to a slower temperature rise under transient forcing than a short-memory response (Rypdal and Rypdal, 2014, Rypdal, 2016). The physical reason is that a long-memory response is associated with energy transport from the surface into the abyss and hence slower temperature rise at the surface. Hence, if the GCMs contain a response on even longer time scales than the long scale in the two-box model the result would be a slower temperature rise in the GCMs than in the two-box model for the 1pctCO2 forcing.

As we understand Merlis et al. (2014), volcanic forcing and abrupt CO2 change yield similar values for the fast component of the climate sensitivity in GCMs, but 5-15% smaller than the transient climate sensitivity. For the same reason as explained above, long memory in the response will give rise to a lower transient response and an underestimation of the sensitivity. Hence, these effects do not necessarily imply nonlinearity in the response.

[3] Agree, see e.g., Fig. 8 in Rypdal and Rypdal (2014). We have decided to omit the reference to Andrews et al. and include several others that are more relevant.

[4] For global GCMs we know that a two-exponential, or a power-law, response function work quite well, and we have a pretty good ide why. It has to do with the different thermal inertias of the mixed layer and the deep ocean, and the rate of heat exchange between the two. We have much less clear ideas about the response function of the ZC-model. As Reviewer #2 pointed out. The ZC model is very different from a GCM. The 25 yr time delay response to the slow solar forcing is solely based on visual inspection of the forcing and response time series is admittedly very crude, but we have no reason to believe that a more sophisticated response function is any better. Since the purpose here is just to find an estimate of the variance of the internal variability, we think the approach makes sense.

[5] Along the same lines. The reason why we cannot do this in a meaningful way is that we have so poor knowledge about the response function for the ZC model on the short time scales. The reason we chose a harmonic oscillator model in Fig. 4 is the apparent enhanced ENSO oscillations after major volcanic eruptions. If we use a certain response function and we get different fits for the sum of responses from the responses to the sum of forcing we can always blame the incorrect response function. Hence, this will not construe another test.

[6] In principle we agree that we should put confidence intervals on these two curves to demonstrate that they are not significantly different from each other. This could easily be done as we do in Fig. 2 of the revised paper by Monte Carlo simulation of 1/f processes. However, in Fig. 1d the two curves to compare are so much on top of each other that they cross each other several times. We have explained this in the revision.

[7] This point was discussed in our response to point [1], but let us elaborate on it here. The "invalid and completely unnecessary approximation" would be apparent by reading Sect. 3.4 in the L&V paper. The approximation is the basis of their Eq. (5), which assumes that one can neglect a cross term which is the product of the solar response and the volcanic response on a given time scale Δt. L&V argue that one can do this because solar and volcanic forcing are statistically independent processes. The approximation would have been OK if we had a large ensemble of realisations of solar and volcanic forcing to average over, but in this case we have only one realization of each (the historic forcing over the last millennium). One of us (K. Rypdal) was a reviewer of the paper and pointed out this weakness in the first review. The result was that L&V kept the old results, but added a paragraph at the end of page 8 where they admit that "the cancellation of the cross terms assumed by statistical independence is only approximately valid on single realizations, especially at low frequencies where the statistics are worse."

The source of this error is probably rooted in the sloppy notation of using angular brackets <> for averages which are really not ensemble averages (or expectations) but rather *estimates* in the form of time averages. If two quantities X and Y are statistically independent their the expectation E[XY]=0, but the time-average estimate <XY> is normally nonzero, and on long time scales Δt we have have virtually no statistics, so there is no reason to believe that <XY> is a good estimate of E[XY]=0.

To us it seems clear that L&V have understood the error, and the appropriate response would be to omit this approximation and replace the blue curve in their Fig. 3b with the one computed without this approximation. But then this Fig. 3b would look similar to our Fig. 2, and obviously be much less convincing. Instead they present the "correct" curve as a ratio given by the lower curve in their Fig. 3c, along with the "incorrect" ratio (the upper curve). The "correct" ratio is probably more or less the same as we would get if we compute the ratio between the red and the blue curve in our Fig. 2 (our results are not completely identical to L&V, which may be due to slightly different steps between the values of Δt where the Haar fluctuation is computed – but we have used codes downloaded from Shaun Lovejoy's web site). We also find that the red curve is higher than the blue for 200< Δt <1000 yr, but on these time scales the fluctuation level is estimated from 5 effective data points for Δt=200 yr, and for only 1 effective data point for Δt=1000 yr. Actually the number of effective data points is even smaller because the time series is not white noise, but exhibits dependence on all scales. Hence, it is obvious that these differences are not statistically significant.

So there are two issues here. One is scientific; the results without the approximation are not significant. The other is the way L&V are presenting their results. In the revision we have decided not to dwell too much on L&V's presentation and focus on the results.

[8] We included the reference to MacMynowski et al., Geoffroy et al., and Fredriksen et al., which have presented spectra for a large number of CMIP5 models.

[9] One should keep in mind here that the harmonic oscillator response was employed for comparison with the ZC model, which responds very differently from GCMs. As stated in the paper, the purpose of this demonstration was not to present a realistic response model for either of the model results analysed by L&V. It was simply to demonstrate that the effects which L&V attributes to nonlinearity is easily produced in linear response models with internal noise. And as a pedagogical tool, we think a driven, damped harmonic oscillator is an excellent choice.