Patrick Frank 12 April 2015

On the reliability of global air temperature projections in light of propagated error: A critical review

Manuscript # EARTH-S-15-00044

Response to Reviewer #1

Summary

This reviewer:

- never grasped the central argument that linear extrapolation of GHG forcing is vulnerable to linear propagation of error;
- ignored so much relevant analysis that the manuscript must have been read extremely carelessly, or was not fully read at all; AR items 1.1, 1.1.1, 1.2, 3.3, 5.3.2, 5.4, 6.2.2, 6.3.2, 6.4, 11.1, 14, 15;
- diagnosed fundamental problems that were neither fundamental nor problems; AR items 5.1, 6.2.2, 7.2, 14;
- misconceived eqn. 6 as a climate model, resulting in multiple errors of critical review, AR items 1.1, 3.2, 3.3, 4.1, 6.5.1, 17.2;
- shows no understanding of the meaning of uncertainty, or its distinction from physical error, AR items 1.2, 5.3, 5.4, 7.2;
- claims a constant model error that has no literature basis, AR items 6.3.2, 6.4;
- confused the mean of ensemble errors with the error in a mean state, AR item 7.1;
- invariably confused accuracy and precision, AR items 11.1, 13.1, 13.2;
- shows no familiarity at all with the meaning, the method, or the display of propagated error, AR items 5.4, 11.2, 13.1-13.4 and has missed no opportunity to confuse error and uncertainty.

Reviewer comments are presented in full, indented, numbered, and in italics. Each separate reviewer idea is sectioned out. Author responses follow.

1. *The submitted paper attempts to document the impact of structural errors in cloud radiative forcing within the current generation of of GCMs on the long term climate projections made by those models. The analysis is conducted by proposing a linear model relating global mean forcing of the climate system to temperature response.* **[Author Response (AR) item 1.1]** *The author proposes that the systematic error in radiative balance the CMIP5 generation GCMs introduces a potential error* **[AR item 1.2]** *orders of magnitude greater than the range of temperature projections seen in the 5th assessment of the IPCC, and thus concludes that the GCMs are not a reliable source of information to inform future climate projections.*

1.1 Manuscript eqn. 6 does not relate, "*global mean forcing of the climate system to temperature response*." Rather, eqn. 6 relates simulated forcing in general circulation models (GCMs) to GCM air temperature projections. That is, the model pertains to model behavior, not to the behavior of the climate.

This point is absolutely critical. However, the reviewer apparently missed this idea despite that it is plainly stated at the outset, e.g., abstract sentence 3 and Introduction paragraph 2.

1.1.1 The eqn. 6 linear emulation model is not "*proposed*." Rather, it is demonstrated. GCM air temperature projections are demonstrated to be linear extrapolations of GHG forcing; e.g., manuscript Figures 2, 3 & 4. Auxiliary Material (AM) Figures S1, S3-S7, and S10. The reviewer never addressed these demonstrations.

1.2 Propagated systematic error produces large uncertainties, not large errors.

Uncertainty represents ignorance, not error: page 31 line 5, manuscript sections 2.4.2, 2.4.3, and 3. Auxiliary Information Section 10.2 provides a detailed discussion of the meaning of uncertainty.

However, the review proceeds as though the reviewer had never read any of this. Thus, the reviewer's analytical summary is not correct.

2. *The author motivates his study by considering the lack of end-to-end uncertainty analyses in GCM projections, and to some degree, he is correct that a comprehensive study documenting the propagated uncertainties of the assumptions made in GCM parameterizations and systematic choices* **[AR item 2]** *would be desirable - and although various studies have made significant inroads into making such an assessment, the field is far from complete.*

2. The manuscript is not at all concerned with, "*assumptions made in GCM parameterizations and systematic choices*." The manuscript concerns true physical accuracy. The known, i.e., **<u>observed</u>**, empirical average physical error of CMIP5 GCM projections is described and propagated. The reviewer's understanding is again incorrect. It is not about assumptions and choices, but about observational error.

3. *However, this author's proposed error propagation study is certainly not an accurate assessment of model error* **[AR item 3.1]**, *and confuses a number of fundamental properties of the climate system* **[AR item 3.2]** *and how it might behave in a zero-order model.* **[AR item 3.3]**

3.1 The ±4 Wm^{-2} average long-wave cloud forcing error from (Lauer and Hamilton,

2013) is certainly an accurate assessment of model error.

3.2 The analysis does not concern the climate system, but concerns GCM projections. The reviewer's misunderstanding of this point induces a pervasive misconstrual throughout the review. See below.

3.3 Manuscript eqn. 6 is not a "*zero-order model.*" It emulates GCM behavior. It has no relevance to climate.

This point is so crucial, that it is made immediately in introductory paragraph 2. Nevertheless, the reviewer missed it.

The focus on GCM emulation (as opposed to climate) is repeated on page 7, paragraph 1; page 15, paragraph 1, and; page 16, line 8. Figures 2, 3 and 4 specifically reference eqn. 6 to GCM output, not to climate observables.

Page 19, line 1 specifically notes that, "*These examples empirically validate the PWM as an accurate emulator of GCM global average air temperature projections.*

This distinction is absolutely central and is made again here: Eqn. 6 demonstrates that GCMs just linearly extrapolate GHG forcing to project global air temperature.

Eqn. 6 is stated in the manuscript to have no necessary connection to the physical climate. It is not a climate model. It is never represented as a climate model. It should never be understood as a climate model.

There is no ambiguity: eqn. 6 is a GCM emulator. It is not a simple climate model.

In missing this critical point, the reviewer's entire critique was misconceived from the outset.

The author adds that climate modeler reviewers have consistently confused this point.

4. *Firstly, the linear emulator proposed by the author represents the global mean temperature as a linear function of the forcing.* **[AR item 4.1]** *This effectively ignores any thermal inertia in the system, and assumes that the system will instantly equilibrate to a new forcing level. Ignoring the thermal mass of the ocean is not an appropriate assumption, even for a simple emulator.* **[AR item 4.2]**

4.1 The reviewer's view is not correct. The PWM represents **<u>GCM air temperature projections</u>** as a linear function of forcing. The PWM emulates the behavior of GCMs. It has nothing whatever to do with global climate (AR item 3.3). Linear GCM behavior is fully demonstrated (manuscript Figures 2, 3 & 4, AM Figures S1, S3 - S7, and S10)

4.2 The PWM (eqn. 6) successfully emulates the global air temperature projections of GCMs. That is all. This success demonstrates that GCMs project global air temperatures as linearly extrapolated GHG forcing. The reviewer has passed over this demonstration in silence.

The PWM is not a model of climate (AR item 3.3). Thus, the reviewer's criticism is irrelevant.

The reviewer went on to ignore the success of the PWM, and overlook the consequent validity of linear propagation of error.

5. *However, even allowing for this oversight,* **[AR item 5.1]** *the critical flaw in the author's logic is the treatment of errors in global mean cloud forcing, and how this is then related to each year's incremental forcing change. The author proposes that there is a +/- 4Wm^-2 error in global mean longwave cloud forcing in the CMIP5 generation GCMs.* **[AR item 5.2]** *He then proposes, as a result, that this error accumulates over time,* **[AR item 5.3]** *i.e. that year 1 is subject to a 4Wm^-2 uncertainty, year 2 would be almost 8Wm^-2 etc.* **[AR item 5.4]**

5.1 The judgment of "*oversight*" regarding ocean thermal inertia rests on the reviewer's incorrect supposition that manuscript eqn. 6 is a simple climate model. See AR items 1.1, 3.3, 4.1 and 4.2.

5.2 (Lauer and Hamilton, 2013) reported the ±4 Wm^{-2} CMIP5 rms average annual observed long wave cloud forcing (LWCF) error. The author did not "*propose*" it.

5.3.1 Accumulation of uncertainty is not an author proposition. It necessarily follows from the linearity of GCM projections; cf. manuscript eqns. 1 and 2. Theory bias error is a property of the model; it is thus necessarily present in every single projection step.

5.3.2 Uncertainty, not error, accumulates with each theory-bias-burdened calculational step. The ±4 Wm^{-2} LWCF error is from a systematic theory bias (ms section 2.3.1) of climate models. Uncertainty increases because model theory-bias error is present in every model projection step.

This is discussed in detail in ms. section 2.4.2, but the reviewer apparently missed it.

5.4 The annual error is never larger than ±4 Wm^{-2}. The reviewer is not correct. Uncertainty increases, not error.

However, eqn. 7 as written requires reader understanding of propagated error. Absent this, one can see how it might cause the reviewer's misperception. Therefore, eqn. 7 and text has been slightly revised to clarify this meaning.

Manuscript section 2.4.3, especially p. 33, par. 2ff, details the distinction between error and uncertainty. The reviewer gives no evidence of having read this section, nor AM page 26, par. 2ff, where the distinction is discussed further.

Unfortunately, review comment 5 again shows that the reviewer does not understand error propagation.

The uncertainty envelope describes the increasing lack of knowledge concerning the position of the projected climate in its simulated phase-space relative to the phase-space position of the physically true future climate. That lack of knowledge becomes increasingly worse as the number of simulation steps increases, because of the unceasing injection and projection of theory-bias error.

The uncertainty grows without bound, because it is not a physical quantity. It is an ignorance width. When the width becomes very large, the simulation no longer has any knowable information about the physically true climate state.

These last two paragraphs have been added into manuscript page 40.

6. *In making this assumption,* **[AR item 6.1]** *the author is confusing forcing and feedback.* **[AR item 6.2]** *It is true that the models in AR5 exhibit systematic errors in cloud forcing. This is well documented in the very papers the author uses to obtain his cloud forcing data. No author has suggested that these errors are random, or should be treated as such. But the biases are just that, they are constant,* **[AR item 6.3]** *and each model has already equilibrated to whatever mean state cloud bias that it exhibits.* **[AR item 6.4]** *In the author's simple linear model without an ocean (eq. 6), the +/-4Wm^-2 should be appended to the F_0 term, not to the \Delta F_i term.* **[AR item 6.5]**

6.1 There is no assumption of error growth, AR items 5.3 and 5.4.

6.2.1 Forcing and feedback are correctly defined, i.e., manuscript p. 6, par. 3, and p. 7, section 2.1.

6.2.2 There is no confusion about forcing and feedback. The reviewer comment neglects the point actually at issue, which is how well the tropospheric thermal flux bath is known at each step in any simulation. See below and also item 6.5.2.

LWCF error puts uncertainty into the magnitude of the tropospheric thermal energy flux bath. Both long wave cloud feedback and GHG forcing contribute to the total thermal energy flux of the troposphere.

CMIP5 LWCF error means that the thermal energy flux of the troposphere is not simulated to an annual average resolution of better than ±4 Wm^{-2}. This defines a lower limit of CMIP5 model resolution. Therefore, the climatological impact of GHG forcing cannot be simulated with more annual accuracy than ±4 Wm^{-2}.

The annual average GHG forcing increase of ~0.035 Wm^{-2} in is 114× smaller than the lower limit of CMIP5 model resolution of tropospheric thermal energy flux, and therefore its effect cannot itself be resolved in a simulation.

The reviewer apparently did not read section 2.4.1, ms p. 27, where this point is made.

The reviewer apparently also does not grasp that the projected change in tropospheric air temperature derives from a mergence of GHG forcing into the tropospheric thermal energy flux bath.

In short, if the model error in cloud feedback is large, the tropospheric thermal flux bath is not simulated to sufficient accuracy, and thus the impact of a small perturbation -- GHG forcing -- cannot be resolved.

6.3.1 The ±4 Wm-2 error is an average CMIP5 bias, not a constant bias.

6.3.2 The reviewer declares that model errors are constant and can be removed by differencing, but did not support this assertion with any reference to the published literature. A reason may be that there is no published literature to reference.

Manuscript section 2.4.3, p. 33ff, and AM section 7 discussed this assumption in detail. But once again, the reviewer apparently failed to read it.

From the Auxiliary Material Section 7.1.1: "current advanced climatology and climate modeling texts do not mention removal of systematic model error by differencing against a base-climate.

"Nor is linear response theory invoked therein to justify error-removal by differencing. Neither Chapter 9, “Evaluation of Climate Models,” nor the Technical Summary of the 5th Assessment Report of the IPCC mention removal of model systematic error by differencing against a base-climate simulation.

"At this writing there appears to be no body of published literature establishing the linearity of climate model error by comparing sequences of hindcast climates with the observed evolution of climate. That is, invocation of linear response theory to justify removal of error by differencing has apparently not been tested and validated against the physically real climate."

That is, there are no published grounds to substantiate the reviewer's claim of constant error.

6.4 Model error is discussed in ms. section 2.4.3 and AM Section 7.1. There is no published precedent for the reviewer's claim that simulation error is constant for any model.

Further, the reviewer apparently does not realize that errors cannot be constant when arising from theory bias. The reason is that, with theory-bias, every climate state is incorrectly simulated. In a step-wise projection, the prior erroneous prior climate state is projected incorrectly. When state errors are incorrectly projected, their magnitudes are modified by theory bias. They cannot be assumed to proceed unchanged. They must typically take different values in the subsequent state.

6.5.1 Eqn. 6 is not a "*simple model*" of climate; *cf.* AR items 1.1, 3.3, 4.1, and 4.2.

6.5.2 The logic of appending ±4 Wm-2 to the ΔF_i is itemized for reviewer understanding:

1. The ±4 Wm^{-2} of LWCF error is the average lower limit of CMIP5 model resolution of the simulated full thermal flux bath of the troposphere.
2. In each simulation step "*i*," air temperature is determined by the total tropospheric thermal energy flux, F_i Wm^{-2}.
3. As the ±4 Wm^{-2} derives from theory-bias, it is inherent in every simulation step.
4. At every simulation step "*i*" in a GHG projection, the initial simulated tropospheric flux bath is $(F_{i-1} + \Delta F_i)$±4 Wm^{-2}, where ΔF_i is the increase in GHG forcing.
5. Thus, at every simulation step "*i*," the simulated tropospheric thermal energy flux is F_i±4 Wm^{-2}.
6. The ±4 Wm^{-2}, persistent in every simulation step "*i*," necessarily conditions the simulated impact of every ΔF_i.
7. The simulated impact of each ΔF_i on air temperature cannot be resolved to better than the lower limit ±4 Wm^{-2} of model resolution of the tropospheric thermal energy flux bath.
8. The annual ±4 Wm^{-2} resolution limit from theory-bias is much larger than the annual ΔF_i.
9. Therefore, the air temperature impact of the ΔF_i is below the model resolution limit.
10. As ±4 Wm^{-2} >> ΔF_i, propagated uncertainty will always exceed ΔT_i.
11. Because of the persistent ±4 Wm^{-2} theory-bias, every simulated F_i in a futures projection is incorrect by some unknown magnitude.
12. Hence, from 11, the propagation of error and the step-wise increase in uncertainty.
13. This analysis was presented in detail in manuscript Section 2.4.3.
14. The physical reasoning should be obvious to any physical scientist.

Ironically for the reviewer's argument that, "*the +/-4Wm^-2 should be appended to the F_0 term*," in a simulation the F_0 can be specified, and therefore need not be conditioned at all by model theory-bias.

7. What the author should be considering in this study is the GCM uncertainty in cloud *feedback. For sure, the mean state bias* **[AR item 7.1]** *can inform the degree to which we trust each GCM. Various studies (Sherwood et al 2014, Fasullo and Trentberth 2013 amongst others) have documented promising methodologies for relating the mean state biases* **[AR item 7.1]** *to feedbacks, but this remains an active field of research. But the author's approach, assuming that the bias will accumulate with each year of the simulation is simply incorrect.* **[AR item 7.2]**

7.1 The ±4 Wm^{-2} is not a mean state bias. It is the mean bias (error) of an ensemble of simulated states. They are not the same at all; shown below.

The mean bias (error) of an ensemble of model simulations is, $\pm\bar{\sigma}_N = \sqrt{\frac{1}{(N-1)}\sum_{i=1}^{N}\sigma_i^2}$,

i.e., the root-mean-square (rms) of the expectation value errors of *N* individual modeled states. This is the ±4 Wm^{-2} of (Lauer and Hamilton, 2013).

In contrast, the error in a modeled mean state is, $\pm\sigma_m = \sqrt{\sum_{i=1}^{n}\sigma_i^2}$, i.e., the root-sum-square (rss) of the sequential errors made by a given model in the "*n*" steps of calculation to the final state extrapolated from the mean set of initial conditions.

The first is the mean of the errors produced by *N* individual models when calculating a given state; a multi-model mean. This is the cloud forcing error presented in (Lauer and Hamilton, 2013).

The second is the final uncertainty obtained by propagating the errors made by one individual model through the sequence of *n* calculations to the final state of the mean state of any ensemble set of states. The two errors are not the same at all.

7.2 There is no assumption that bias accumulates. Bias error propagates; uncertainly accumulates. The reviewer has consistently displayed this confusion between error and uncertainty, and plainly does not understand error propagation. Propagation of a consistently repeated error, such as theory-bias, absolutely requires increased uncertainty with every calculational step. This requirement is explicit in eqns. 1 and 2.

Manuscript page 41 quotes the JCGM Guide to Expression of Uncertainty exactly to the effect, (JCGM, 100:2008) with additional references to US NIST and other published literature (Roy and Oberkampf, 2011; Taylor and Kuyatt., 1994), that uncertainty accumulates when error is propagated through sequential calculations.

The reviewer's judgment is contradicted by the exactly relevant and authoritative literature; literature that the reviewer has completely ignored.

8. On the basis of these fundamental errors in the logic of the study, **[AR item 8]** *I find the paper unacceptable for publication.*

8. The reviewer consistently misinterpreted eqn. 6 as a simple climate model, rather than as a GCM emulator, and thus produced a completely misconceived review.

The reviewer has also consistently confused error and uncertainty, and displayed no understanding of the uncertainty arising from propagated error.

The reviewer has displayed no understanding of error propagation itself, or of why error propagation is required of a set of sequential calculations, or of the fact that error propagation has been standard in the physical sciences for far more than 100 years.

The reviewer displayed no understanding that theory bias is inherent to the model itself, and so must impact every single step of a simulation.

The reviewer has made claims that model errors are constant and can be differenced away, which are unsubstantiated in the published literature, including the IPCC AR5.

The reviewer clearly did not carefully read the manuscript or the AM, given the many lacunae in the reviewer's argument.

For these reasons, and others noted above, the reviewer's judgment to reject has no scientific merit.

Smaller review items:

In addition, I note the following additional minor points.

9. Page 3, 1st para: The likelihood of the future climate warming significantly is not conditional only on the use of GCMs. **[AR item 9.1]** *Considerations of the paleoclimate record, the observed warming during the satellite era alone or simply the radiative impact of increasing CO2 alone (as the author presents) makes it hard to make the argument that the Earth will not warm in a high CO2 future.* **[AR item 9.2]**

9.1 The reviewer seems to not know that in science the meaning of data is established only by reference to the unique prediction of a valid and explanatory physical theory. The current physical theory of climate is present only in GCMs. Therefore, only GCMs can give meaning to surmises of future warming from GHG emissions.

Theory-vacant surmises about the meaning of the paleoclimate record or of recent warming are merely tendentious opinion.

9.2 The author expressed no views in the ms concerning the actual impact of increasing GHGs on the terrestrial climate.

The author's conclusion is that the large uncertainties from propagated error reveal that GCMs are unable to reveal the impact, if any, of increased GHGs.

10. Page 3, 2nd para: A skillful representation of atmospheric processes is exactly what gives us confidence that GCMs are more meaningful than a simple linear extrapolation of global mean surface temperatures. **[AR item 10.1]** *Their ability to represent complex coupled processes: ENSO, the Madden Julian Oscillation, the climate of the deep past - is exactly what gives us confidence.* ***[AR item 10.2]*** *These processes are emergent properties of the simulated system, conditional on the representation of dynamics, radiation, clouds, ocean currents, sea ice, biological feedbacks. Thus the ability of the model to represent these coupled processes as a sum of simpler parts gives us confidence that the system we have built is able to represent the emergent behavior of the climate.*

10.1 How do the reports of Jiang, et al., (Jiang et al., 2012), Lauer and Hamilton (Lauer and Hamilton, 2013), Su, et al., (Su et al., 2013), among others, which demonstrate large scale errors in simulation of cloud processes, support a claim of "*skillful representations of atmospheric processes*"?

10.1.1 The reviewer has also neglected to mention that "*skillful*" means statistically correlated, not physically accurate. Parameterization schemes that allow GCMs to mimic, e.g., the Madden-Julian Oscillation (MJO) are not known to be more physically accurate than those parameterizations that do not.

10.1.2 Nothing in the reviewer comment gainsays the demonstrated fact that GCM air temperature projections are mere linear extrapolations of GHG forcing.

Nor does the reviewer show how the consequence of linear propagation of error is refuted by these representations. Nor does reviewer comment 10 refute the projection uncertainty envelopes.

In sum, reviewer comment 10 has no critical content.

10.1.3 Unless the representations of climate are known to derive from physically valid theory, they have no unique physical meaning. The reviewer's confidence merely reiterates the consistent failure to distinguish between model precision and physical accuracy.

The reviewer has also implicitly reinforced the point made in the introduction, namely that model simulations are typically never conditioned with physically valid error bars.

10.2 Additionally, the IPCC itself contradicts the reviewer claims of skill:

10.2.1 ENSO (IPCC AR5 Section 9.5.3.4.1, p. 908ff): "[A]s was the case in the AR4, simulations of both background climate ... and internal variability **exhibit serious systematic errors**, many of which can be traced to the representation of deep convection, trade wind strength and cloud feedbacks, with little improvement from CMIP3 to CMIP5. ... **The CMIP5 models still exhibit errors in ENSO amplitude, period, irregularity, skewness, spatial patterns or teleconnections.**" (author bold)

10.2.2 MJO (IPCC AR5 section 9.5.2.3, p. 796ff): "**[M]odel errors** in representing the MJO [include] convection parameterizations [that] do not provide sufficient build-up of moisture in the atmosphere for the large scale organized convection to occur [and b]iases in the model mean state. ... High-frequency coupling with the ocean is also an important factor. ... **[S]imulation of the MJO is still a challenge for climate models.** Most models have weak coherence in their MJO propagation (smaller maximum positive correlation)." (author bold)

These model errors indicate that large errors exist in the simulated partitioning of energy among the climate sub-systems; certainly much larger than the annual 0.035 Wm^{-2} increase in CO_2 forcing and even than the $\pm 4Wm^{-2}$ long-wave cloud forcing error.

Not only are the reviewer's confident assurances misplaced, but the given examples demonstrate the opposite of what the reviewer intended, namely that GCMs are unable to resolve the impact of GHGs on tropospheric temperature.

One also wonders at the reviewer's confidence in ocean models that do not converge. (Wunsch, 2002)

11. Page 2, 1st para: Parameter sensitivity tests are not tests of precision. **[AR item 11.1]** *They are tests of propagated error in the purest sense.* **[AR item 11.2]** *Our assumptions are the parameter values, and by varying these parameters within a range of plausibility and running climate simulations we can assess how these assumptions are impacting our projections of long term climate change.* **[AR item 11.3]**

11.1 The reviewer has completely ignored the standard definitions of precision and accuracy provided in the Introduction, and is not correct about sensitivity tests and precision. The provided definitions are referenced to The International Guide to Metrology and to US NIST; completely authoritative sources.

Further, from (Bevington and Robinson, 2003): "*The accuracy of an experiment is a measure of how close the result of the experiment is to the true value; the precision is a measure of how well the result has been determined, without reference to its agreement with the true value. The precision is also a measure of the reproducibility of the result in a given experiment.*"

Parameter sensitivity tests merely tabulate model internal variability, which is identical to precision. They reveal nothing of accuracy; the difference relative to observations.

Parameter tests also show that models are incapable of producing a unique result, meaning that accuracy cannot be determined in any case.

11.2 The reviewer again has demonstrated confusion about the meaning of error propagation. Eqns. 1 and 2 rigorously define error propagation. Parameter tests do not follow these equations. Therefore, parameter tests cannot be error propagation in any sense, no matter the reviewer's insistence.

11.3 The reviewer has nicely defined a test of model precision, i.e., varying assumptions to see how model output varies *without reference to any true value* (item 11.1).

12. Page 2, 1st para: Taylor diagrams, however are not a measure of accuracy **[AR item 12.1]** *because they are not predicting an out-of-sample metric.* **[AR item 12.2]** *Taylor diagrams are used to tune climate models to the observed climate. Therefore, they are not measures of predictive skill,* **[AR item 12.3]** *but the degree to which the model has been tuned to replicate the observed climate. Accuracy in future projections is not guaranteed by the models' ability to match the observed climate (although the latter is a necessary condition for us to consider the model a plausible candidate).* **[AR item 12.4]**

12.1 Figure 3 of (Lauer and Hamilton, 2013): the Taylor diagram is observation minus prediction; a standard accuracy metric.

Figure 3 Legend: "Taylor diagrams showing the 20-yr annual average performance of the (top) CMIP3 and the (bottom) CMIP5 models for (left to right) CA, LWP, and ToA SCF and LCF **as compared to satellite observations**." (author bold).

(Collins et al., 2011) also specifically cite Taylor diagrams as an error metric: "There are a number of simple and widely used metrics which may be used to quantitatively compare models with observed climate variables (e.g. (Taylor, 2001)).

Thus, the reviewer is factually incorrect.

12.2 Accuracy need not refer to "*an out of sample metric*." Note the definition of accuracy in item 11.1 from (Bevington and Robinson, 2003). Accuracy refers only to the disparity between a prediction and an observed result.

12.3 The disparity between model hindcasts and the observed climate priors is an obvious and standard measure of accuracy. The reviewer's view is mistaken.

12.4 The reviewer apparently does not realize that 'plausibility' is not a standard of performance in the physical sciences.

13. Page 6, 2nd para: "error bars" are regularly published in studies of propagated errors - see Sexton et al (2013), **[AR item 13.1]** *Rowlands et al (2012),* **[AR item 13.2]** *Collins (2012),* **[AR item 13.3]** *Yamakazi (2013)* **[AR item 13.4]** *amongst many others.*

13. Not one of the reviewer examples of propagated error actually includes propagated error; shown below.

13.1 (Sexton et al., 2012) (not 2013) nowhere mention error propagation. Nowhere is error propagated through a projection. Nowhere is shown the increasing uncertainty necessarily inherent in a theory-biased step-wise simulation.

Sexton, et al., instead derive a "discrepancy" term to represent model error. Presumably, this is what the reviewer has in mind. Discrepancy is, "*based on the assumption that structural differences between HadSM3 and other models are a plausible proxy for the uncertain effects of structural errors in how HadSM3 represents climate processes in the real world.*" (Sexton, p. 2526)

That is, physical error, i.e., accuracy, is presumed to be identical to inter-model variance, i.e., precision. Inter-model variance is a form of precision, not of accuracy.

Further, nowhere is discrepancy propagated through a futures simulation, as eqn. 2.

13.2 (Rowlands et al., 2012) write, "*Here we present results from a multi-thousand-member perturbed-physics ensemble of transient coupled atmosphere–ocean general circulation model simulations."* (Abstract, p. 256)

and go on to state that, *"Perturbed-physics ensembles offer a systematic approach to quantify uncertainty in models of the climate system response to external forcing, albeit within a given model structure.*" (p. 256)

That is, Rowlands, et al., represent error as model variance: precision, not accuracy.

13.3 In (Collins et al., 2011) (not 2012) propagation of error is never mentioned. Collins Figure 3 shows physical uncertainty as model variability about an ensemble mean; i.e., a measure of precision.

Although error is calculated as the difference between observations and a bias-removed model mean, this error is never propagated through a climate simulation.

13.4 In (Yamazaki et al., 2013), once again, propagation of error is nowhere mentioned. "*Prediction error*" is defined as "*predicted values from the random forest* [from their random forecast statistical algorithm] *against simulated values...*" (p. 2786). Thus, their errors are the differences between test simulations and their targeted assemblage of standard simulations. That is, error is again model variance, a precision metric.

Yamazaki, et al., Figure 9 also reports air temperature errors differences between simulations and observables, but those errors are never propagated through a simulation.

In short, not one of the reviewer's examples of propagated error in fact includes propagated error. One can only conclude that the reviewer has no concept of error propagation.

14. Page 6, 2nd para: systematic energy flux errors are not inputs to the system, they are resolved outputs (which might exhibit errors). This is the origin of the author's primary logical miscalculation. **[AR item 14]**

14. In any step-wise simulation, the output of each prior step, *i-1,* constitutes the input of the subsequent step, *i*. It is little short of incredible that the reviewer does not seem to realize this. Or its impact.

Section 2.4.2 introduced the idea of sequential error in a step-wise simulation. Section 2.4.3 discussed this idea in considerable detail. However, the reviewer did not address this discussion, and gives no indication of ever having read it.

The reviewer's "*origin of the author's primary logical miscalculation,*" is instead a reviewer misunderstanding of the very basics of uncertainty in a step-wise calculation; most especially one featuring theory-bias error.

15. Page 7, 2nd para: evaluating total cloud fraction is a complex process, and the different GCMs report it in different ways. It is an entire field of study to assess how best to compare observed cloud properties to their representation in GCMs, requiring satellite simulators to be built into the GCMs themselves to replicate the inverse process which is used to detect cloud properties from satellites. This field cannot be realistically summarized with general statements about "Global" cloud fraction without strictly defining how cloud fraction is to be defined. **[AR item 15]**

15. Global cloud fraction was taken from (Jiang et al., 2012), and was discussed in detail on ms. section 2.4, page 21ff, and in SI Section 6, p. 11ff. However, the reviewer gives no indication of having read this discussion.

16. Page 8, 1st para: what is the justification for ignoring all other feedbacks rather *than the water vapor feedback? Longwave and (primarily) shortwave cloud feedback is our primary uncertainty in future climate change, but any comprehensive feedback model also needs to consider land ice, sea ice and land surface feedbacks, not to mention carbon feedbacks and ocean circulation feedbacks*. **[AR item 16]**

16. The justification for the restriction is obvious and immediately precedes page 8 (Section 2.1). Page 6, par. 3 notes: "The approach begins by evaluating the onset of CO_2-induced greenhouse (GH) warming. The fractional contribution for the water-vapor enhanced (wve) greenhouse effect of CO_2 is thereby assessed, *as applicable to GCMs*." (original italics)

Why should an analysis restricted to an appraisal of the "the onset of CO_2-induced greenhouse (GH) warming," include any other feedbacks? This analysis is further justified by (Manabe and Wetherald, 1967) on whose evaluation of CO_2 forcing the analysis is based. This focus is clearly stated. It is not clear how the reviewer misconceived it.

17. Page 8-9: this section is entirely irrelevant, given that 1ppm CO2 is entirely outside of any Earth-like state. **[AR item 17]**

17.1 The analytical rationale is obvious and immediately precedes page 8 (Section 2.1). Page 6, par. 3 notes: "The approach begins by evaluating the onset of CO_2-induced greenhouse (GH) warming."

How should the onset of CO_2-induced greenhouse (GH) warming be investigated except by deriving the condition of negligible CO_2 forcing? The analytical rationale is patent, and is clearly stated. The reviewer's objection evidences a striking unfamiliarity with physical analysis.

17.2 Further, the derivation is plainly stated to be about CO_2 forcing <u>as it is represented within climate models</u>; not about any Earth-like state.

It is remarkable that the reviewer gives no evidence of ever having realized that the focus is on the behavior of climate models, not of climate, despite this distinction having been emphasized repeatedly in the manuscript.

Reviewer references:

Yamazaki, Kuniko, et al. "Obtaining diverse behaviors in a climate model without the use of flux adjustments." Journal of Geophysical Research: Atmospheres 118.7 (2013): 2781-2793.

Sexton, David MH, et al. "Multivariate probabilistic projections using imperfect climate models part I: outline of methodology." Climate dynamics 38.11-12 (2012):

2513-2542.

Rowlands, Daniel J., et al. "Broad range of 2050 warming from an observationally constrained large climate model ensemble." Nature Geoscience5.4 (2012): 256-260.

Collins, Matthew, et al. "Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles." Climate Dynamics 36.9-10 (2011): 1737-1766.

Author references:

Bevington, P.R. and Robinson, D.K., 2003. Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill, Boston, 320 pp.

Collins, M., Booth, B.B., Bhaskaran, B., Harris, G., Murphy, J., Sexton, D.H. and Webb, M., 2011. Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. Climate Dynamics, 36(9-10): 1737-1766.

JCGM, 100:2008. Evaluation of measurement data — Guide to the expression of uncertainty in measurement Document produced by Working Group 1 of the Joint Committee for Guides in Metrology (JCGM/WG 1). Bureau International des Poids et Mesures, Sevres, France.

Jiang, J.H., Su, H., Zhai, C., Perun, V.S., Del Genio, A., Nazarenko, L.S., Donner, L.J., Horowitz, L., Seman, C., Cole, J., Gettelman, A., Ringer, M.A., Rotstayn, L., Jeffrey, S., Wu, T., Brient, F., Dufresne, J.-L., Kawai, H., Koshiro, T., Watanabe, M., LÉcuyer, T.S., Volodin, E.M., Iversen, T., Drange, H., Mesquita, M.D.S., Read, W.G., Waters, J.W., Tian, B., Teixeira, J. and Stephens, G.L., 2012. Evaluation of cloud and water vapor simulations in CMIP5 climate models using NASA "A-Train" satellite observations. J. Geophys. Res., 117(D14): D14105.

Lauer, A. and Hamilton, K., 2013. Simulating Clouds with Global Climate Models: A Comparison of CMIP5 Results with CMIP3 and Satellite Data. J. Climate, 26(11): 3823-3845.

Manabe, S. and Wetherald, R.T., 1967. Thermal Equilibrium of the Atmosphere with a given Distribution of Relative Humidity. J. Atmos. Sci., 24(3): 241-259.

Rowlands, D.J., Frame, D.J., Ackerley, D., Aina, T., Booth, B.B.B., Christensen, C., Collins, M., Faull, N., Forest, C.E., Grandey, B.S., Gryspeerdt, E., Highwood, E.J., Ingram, W.J., Knight, S., Lopez, A., Massey, N., McNamara, F., Meinshausen, N., Piani, C., Rosier, S.M., Sanderson, B.M., Smith, L.A., Stone, D.A., Thurston, M., Yamazaki, K., Hiro Yamazaki, Y. and Allen, M.R., 2012.

Broad range of 2050 warming from an observationally constrained large climate model ensemble. Nature Geosci, 5(4): 256-260.

Roy, C.J. and Oberkampf, W.L., 2011. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. Comput. Methods Appl. Mech. Engineer., 200(25-28): 2131-2144.

Sexton, D.M.H., Murphy, J.M., Collins, M. and Webb, M.J., 2012. Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. Climate Dynam., 38(11-12): 2513-2542.

Su, H., Jiang, J.H., Zhai, C., Perun, V.S., Shen, J.T., Del Genio, A., Nazarenko, L.S., Donner, L.J., Horowitz, L., Seman, C., Morcrette, C., Petch, J., Ringer, M., Cole, J., von Salzen, K., d S. Mesquita, M., Iversen, T., Kristjansson, J.E., Gettelman, A., Rotstayn, L., Jeffrey, S., Dufresne, J.-L., Watanabe, M., Kawai, H., Koshiro, T., Wu, T., Volodin, E.M., L'Ecuyer, T., Teixeira, J. and Stephens, G.L., 2013. Diagnosis of regime-dependent cloud simulation errors in CMIP5 models using “A-Train” satellite observations and reanalysis data. J. Geophys. Res.: Atmospheres, 118(7): 2762-2780.

Taylor, B.N. and Kuyatt., C.E., 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, National Institute of Standards and Technology, Washington, DC.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophs. Res., 106(D7): 7183-7192.

Wunsch, C., 2002. Ocean Observations and the Climate Forecast Problem. International Geophysics, 83: 233-245.

Yamazaki, K., Rowlands, D.J., Aina, T., Blaker, A.T., Bowery, A., Massey, N., Miller, J., Rye, C., Tett, S.F.B., Williamson, D., Yamazaki, Y.H. and Allen, M.R., 2013. Obtaining diverse behaviors in a climate model without the use of flux adjustments. J. Geophs. Res. - Atmos., 118(7): 2781-2793.