Patrick Frank 24 March 2014
Propagation of Error and the Reliability of Global Air Temperature Projections
JGR-Atm submission 2013JD021338

Response to Reviewer #1:

Summary

- The reviewer has repudiated the distinction between accuracy and precision as a "philosophical rant," when in fact the distinction is central to physics.
- The review evidences a lack in understanding of propagated error, item 3.
- The reviewer has mistakenly assumed that differencing between modeled climate observables is identical to differencing between modeled and physically measured climate observables, items 4 and 8.
- The reviewer is apparently unaware that the large measurement uncertainties vitiate attribution and validation, item 5.

Reviewer comments are presented in full, justified, numbered, and in italics.

1. *I have very few detailed comments on this manuscript.*

Too much of this paper consists of philosophical rants (e.g., accuracy vs. precision),...

Response item 1. The distinction between accuracy and precision is basic to all of measurement physics. Accurate measurements are the test of physical theory and the source of scientific knowledge. It is hardly philosophy.

This manuscript focuses on the accuracy of climate models. The Introduction notes that published model evaluations chiefly concern precision rather than accuracy. Establishing the distinction and relaying its fundamental importance, is therefore an obvious sine qua non for the subsequent analysis.

Chapter 7 in Wilks, "Statistical Methods in the Atmospheric Sciences," presents an extensive discussion of the meaning of model accuracy and resolution. [*Wilks*, 1995] Presumably the reviewer sees this, too, as a "philosophical rant."

Reference of theory to potentially falsifying measurement is the bar that fully separates science from philosophy. Accuracy, as opposed to precision, is the standard of measurement quality. Dismissal of concern for accuracy as philosophy exactly rejects the methodological distinction of science from philosophy. The reviewer has provided an inadvertent but very rich irony.

2. *... several pages of basic radiative transfer theory to outline would should take only a few citations.*

As was also the case for submission 1 reviewer #1, the present reviewer has provided no leading citation demonstrating a prior evaluation of the $[CO_2]_{atm}$ necessary for onset of

significant forcing. I have not found mention of this in Ramanathan's papers, nor does it appear in Arrhenius' foundational paper.

Although this reviewer was apparently provided with the author's responses to the submission 1 reviews, the reviewer missed noticing that the author searched several basic texts on climatology looking for, and not finding, mention of the $[CO_2]_{atm}$ necessary for onset of significant forcing in their treatments of basic radiative transfer.

The point of the radiative transfer section is to establish the $[CO_2]_{atm}$ at which onset of climatologically significant forcing occurs. As was the case previously, the present reviewer has apparently missed this point.

> 3. *The bulk of what the author presumably feels is novel here is completely wrong. In particular, the author has not actually shown that errors are propagating in future projections,...*

R3.1 Systematic model error always propagates into futures projections. [*Roy and Oberkampf*, 2011; *Vasquez and Whiting*, 1998; 2005] There is no other way to establish predictive reliability.

The reviewer is evidently proposing that model error might not propagate forward into a simulation. The comment is extraordinary, apparently averring that theory-bias TCF error can disappear from model output when the model is used to make a futures projection.

Apart from magical intervention, it is hard to envision how TCF theory-bias error – a structural characteristic of the model itself -- will not be necessarily present in every GCM simulation step.

R3.2 It impossible to show that TCF error is itself propagating into projections of future climate states because no independent observational data can exist to evaluate nonexistent future states. This is the very reason for propagated confidence intervals.

The confidence intervals obtained by propagated error represent the reliability of the model-projected future state, not the expected magnitude of error itself in the simulation. Indeed, in the absence of observational referents, a simulated climate that includes systematic errors is indistinguishable from a simulated climate that is error-free.

Confidence intervals represent the level of predictive uncertainty stemming from known model error; i.e., the low state of knowledge. Item 3 displays the same lack of understanding concerning confidence intervals and propagated error as evidenced in the other reviews.

The reviewer is recommended to, "*A Complete Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing.*" [*Roy and Oberkampf*, 2011] Roy and Oberkampf examine in detail how to express the predictive uncertainty of non-

linear numerical scientific models. GCMs are examples of such models. Propagation of error through the model is basic to validation and the evaluation of predictions.

Section 4.5 of Roy and Oberkampf discusses uncertainty in model predictions. They, "*extrapolate the uncertainty structure expressed by the validation metric to the application conditions of interest.*" The validation metric is the observable. The condition of interest is the future state. That is, the error between the modeled reference state and the observed reference state is extrapolated into the model predictions where observational data are unavailable.

They go on, "*As is common in scientific computing, no experimental data is available for the application conditions of interest. Then the extrapolated model form uncertainty is included in the prediction of the model at the conditions of interest as an epistemic uncertainty.*"

"*Extrapolated model form uncertainty*" is the propagated theory-bias error that enters into simulated future states. Roy and Oberkampf exactly describe the analytical method applied in the manuscript.

Manuscript equation 6 accurately simulates the air temperature projections of advanced GCMs. Its use for propagation of model error is thus analytically valid.

> 4. *...but misunderstands the distinction between a base−state "forcing" and the uncertainties surrounding total cloud cover/forcing, from the uncertainties in climate change imbalances. The fact that GCMs do not have correct "absolute values" in variables such as TOA radiation balance, global mean temperature, cloud cover, etc is not novel.*

R4.1 The error in cloud forcing means that models are unable to resolve the global annual average thermal state of the troposphere; on average to within ±4 Wm^{-2}. A climate change imbalance, i.e., GHG forcing, is necessarily undetectable in model output when it is two orders of magnitude smaller than a lower limit of model resolution.

The ±4 Wm^{-2} of TCF error is a minimum of CMIP5 climate model error. As a complete accounting of error would show the energy resolution of climate models to be far poorer than this, the physically correct climate response to the imbalance represented by GHG emissions is well beyond the reach of any current simulation.

There is no relief from this situation to be had in model anomalies. See the discussion under item R8.1 below. Manuscript section 2.4.3, which treats this problem explicitly, was not addressed by the reviewer.

R4.2 The reviewer appears to believe that, when discussing the physically real climate, "*uncertainties surrounding total cloud cover/forcing,*" can be somehow removed "*from*

the uncertainties in climate change imbalances." so as to isolate and assess the latter alone.

However, the tropospheric thermal energy flux itself contains no information about its sources. Tropospheric Wm^{-2} arising from cloud forcing are not colored differently from Wm^{-2} entering from GHG forcing. Climate responds coherently to the totality of the tropospheric energy flux, not piece-wise to forcing from this or that source.

When model cloud forcing error is ±4 Wm^{-2}, the modeled thermal flux bath of the entire troposphere is unknown to the amount of ±4 Wm^{-2}. How does the reviewer propose to establish the accuracy of a modeled tropospheric response to a "*climate change imbalance*" that is two orders of magnitude below the resolution of the model?

Specifically, suppose the terrestrial climate were adjusting cloud cover to offset the energy due to GHG emissions such that there were no changes in tropospheric sensible heat. This adjustment would be worth 0.035 Wm^{-2}, annually. With a model TCF annual average resolution of ±4 Wm^{-2}, how would a model be capable of resolving an average 0.035 Wm^{-2} cloud response?

This problem was discussed in manuscript lines 429ff, and especially lines 655-675. However, the reviewer has not addressed it.

R4.3 There is no manuscript claim of novelty concerning model error. The claim of novelty is in the development of a method to propagate error through GCM air temperature projections, and in the results following from the success of this exercise.

> 5. *There is no evidence provided by the author those known issues contaminate our understanding of attribution (which depends on the spatio-temporal evolution of patterns in stratospheric cooling, global OHC increases, etc)...*

R5.1 The propagation of systematic theory-bias model error, which the reviewer wrongly has rejected, provides exactly that evidence.

CO_2-induced cooling of the stratosphere occurs where radiative emission is completely dominant, well outside the regime of the water-vapor-dominated tropospheric climate. It offers no verification of the reliability of tropospheric climate modeling.

R5.2 Evaluation of ocean heat content relies upon Argo buoy measurements. The reviewer may not realize that these buoys have never been field-calibrated. [*Castro et al.*, 2012; *Emery et al.*, 2001; *Hadfield et al.*, 2007; *Xu and Ignatov*, 2013] Therefore, the impact of environmental systematic effects on the accuracy of their measurements is almost entirely unknown. [*Kawai et al.*, 2006] However, inter-comparisons among buoys suggest an rms inaccuracy of ±0.4-0.5 C.

The positive 0-700 m OHC trend as reported by Levitus, 0.4×10^{22} J-yr^{-1}, [*Levitus et al.*, 2009] equates to an annual oceanic temperature increase of 0.004 C; about two orders of magnitude below the level of accuracy of SST measurements. Even 50×0.004 C = 0.2 C, i.e., the entire purported 1955-2005 temperature increase, is below the accuracy resolution of the SST data. The reported OHC clearly suffers from false precision and provides no evidence for attribution.

The bias corrections in Levitus, 2009 do nothing to reduce measurement uncertainty, because the corrections utilize data from instruments that themselves were never field-calibrated. The complete lack of knowledge concerning the intrinsic systematic *in-situ* XBT and MBT bias magnitudes means that subtracting a bias obtained from an independent data source may actually increase the error in the XBT/MBT record.

> 6. *...or in climate sensitivity (for example, the IPCC AR5 plotted absolute global mean temperature against the equilibrium climate sensitivity of the CMIP5 ensemble (Figure 9.42) and found no correlation between the absolute offsets in temperature and the sensitivity of the models).*

R6.1 The fact that different models display very different climate sensitivities is itself clear evidence for lack of physical understanding. The manuscript analysis does not make any connection between model TCF error and variation in model climate sensitivity, so the direct relevance of the reviewer's comment is obscure.

R6.2 Perhaps the reviewer is making comparison to Auxiliary Material Figure S7, which showed a strong linear correlation between the empirical *wve* CO_2 forcing fraction and the 1990 base global air temperature for the CMIP3 models used for the AR4 SRES simulations.

Figure 9.42 in the 5AR uses data taken from IPCC AR5 Figure 9.8 and Table 5. Figure 9.8a and inset show air temperature anomaly projections from a variety of climate models, each referenced to its own 1960-1990 mean. As these models all reproduce the observed HadCRUT4 trend, and nevertheless display different climate sensitivities, they all must have been tuned using anti-correlated parameters, following Kiehl [*Kiehl*, 2007], in order to match the observations.

It is no surprise, then, that these climate sensitivities do not correlate with their absolute 1960-1990 mean, because with anti-correlated tuning every published model mean will have necessarily been removed from a model mean driven by climate sensitivity alone. Thus there is no relevant comparison between AR5 Figure 9.42 and manuscript Figure S7.

> 7. *There is much further extensive discussion of the model performance and biases in that chapter, which I urge the author to read.*

R7.1 Looking through AR5 Chapter 9, "*Evaluation of Climate Models*," one notes that it is completely silent on the failed perfect model tests reported by Collins and by Boer. [*Boer*, 2000; *Boer and Lambert*, 2008; *Collins*, 2002] Chapter 8 in the 4AR, "*Climate Models and Their Evaluation*," likewise completely ignored these tests. The reviewer can satisfy him/herself on the facts of this matter, and is invited to ponder why the failed perfect model tests were not discussed in AR chapters that purport model evaluation.

AR5 Chapter 10 mentions perfect model tests, but only briefly and under *10.6.1.2 Precipitation Extremes*, rather than in a context of air temperature. The perfect model results are misrepresented as indicating the likelihood of anthropogenic attribution in late 20^{th} century extremes, rather than more correctly as indicating the poor likelihood of attribution even given a perfect climate model.

AR5 Chapter 11 discusses the failed perfect model tests under *11.2.1.1 Predictability Studies*, and then puts the best possible face on the fact that predictability is poor even when models are perfect (FAQ 11.1).

R7.2 To be clear: the poor perfect model test results show that even advanced GCMs cannot correctly partition energy into the climate subsystems at the resolution necessary to reveal the effect of 35 mWm^{-2} annual forcing increase. The confidence intervals produced from propagation of these errors will always be larger than any simulated future GHG effect, because the magnitude of error is much larger than the magnitude of GHG forcing.

R7.3 One also notes that the HadCRUT4 trend in AR5 Figure 9.8 does not display the very significant error bars that follow from known systematic measurement error of air temperature sensors. [*P. Frank*, 2010; *Hubbard and Lin*, 2002; *Hubbard et al.*, 2001; *Lin et al.*, 2005]

> 8. *I have not seen all of the review comments to the previous manuscript, but I was provided with the author responses to those reviews, and was able to see several italicized portions of previous review comments. I think that previous reviewer #1, in particular, already diagnosed many of the problems in this current study. The responses provided by the author are not compelling.*

R8.1 The primary criticism of submission 1 reviewer #1 (S1R1) was a "*confusion of base-state forcing with feedback*." The reviewer elaborated that comment as, "*the overwhelming error in this paper is how this uncertainty in cloud forcing is applied in the future projections made using the empirical linear model. Each GCM starts simulations in ~1850 in an equilibrium state, thus all of the errors in base state cloud forcing are already represented in the global mean temperature in 1850.*"

S1R1 is apparently claiming that differencing removes model errors, because all model cloud forcing errors are already represented in the base state climate. I.e., a constancy of

error exists in the equilibrated 1850 base-state and in subsequent modeled states. Differencing then removes this constant error.

But the error treated in the manuscript is not the difference between a modeled climate and its subsequent manifestation. The error is the difference between a modeled climate and its target observations; an entirely different comparison.

Let me try to make the problem clear. Let the physically real 1850 climate be C_{pr1850}. Suppose the equilibrated modeled 1850 climate plus its error is $C_{m1850} = C_{pr1850}+e_{m1850}$. The error in the equilibrated simulation is then $e_{m1850} = C_{m1850}-C_{pr1850}$. How are the structure and magnitude of this error to be determined? The observational details of the 1850 climate are not known.

The simulated climate for 1851 is then $C_{m1850}+\Delta E_{perturbations}\ (Wm^{-2}) \Rightarrow C_{m1851}$. Then $e_{m1851} = C_{m1851}-C_{pr1851}$. As e_{m1850} is both unknown and is instrumental in producing C_{m1851} with its error (e_{m1851}), how is it determined that in fact $e_{m1850} = e_{m1851}$? The magnitude of e_{m1851} is as non-computable as e_{m1850}: there are few or no known 1851 climate observables.

The physical errors in the baseline 1850 climate are thus unknown and unknowable. So are most of the errors in subsequent simulated climate states with respect to most of the rest of the target climate physical observables. How, then, can it be claimed that differencing model simulations removes physical error?

This further discussion will enter into the revised manuscript.

As a relevant aside, the very large uncertainties in the global average annual air temperature record, [*Emery et al.*, 2001; *P. Frank*, 2010; *Patrick Frank*, 2011; *Hubbard and Lin*, 2002; *Lin et al.*, 2005; *Saur*, 1963] must also and necessarily enter into the uncertainty in the expectation values of any GCM using that record as a physical validation target.

R8.2 Although not specifically mentioned the reviewer may have linear perturbation theory (LPT) in mind in support of removing model error by differencing from a base state. LPT describes a linear response of stochastic dynamical processes to small perturbations. Climate models may be constructed to have this property.

If the surmise is correct, then R1R1 implied that LPT describes the response of the physical climate and thus is a valid physical theory of climate. However, validation is something to be demonstrated, not assumed.

Such a demonstration requires an exercise analogous to the 1850 validation outlined in R8.1.

In fact, LPT validation tests have been carried out, at least partially, in comparisons of model hindcasts vs. physical climate observables. [*Jiang et al.*, 2012; *Klein et al.*, 2013; *Lauer and Hamilton*, 2013; *Williams and Webb*, 2009] In every case, model errors appear

in the C_{myyyy}-C_{pryyyy} differences from hindcasted climates. Further, model differencing anomalies are not of zero error with respect to observational climate anomalies.

The error discussed in the manuscript is model vs. observation. LPT does not necessarily apply, first because LPT has not been demonstrated to be a physically correct description of climate response, and second because such tests as do exist do not validate the LPT prediction of zero anomaly error.

There is no reason to think that model error disappears when a model result is differenced against corresponding climate observables. Nor, indeed, is there any reason to think that model error disappears when differenced against itself.

As noted in R8.1, there is also no reason to think that uncertainty is zero when a base-state model that is an incorrect representation of its energy state is projected using a biased theory.

Manuscript section 2.4.3 discussed this in detail, but was apparently ignored by the reviewer.

It is finally not surprising that the present reviewer would see the R1R1 criticism sympathetically, because the same mistaken assumption of removal of physical error through model differencing is evident in review item 4.

References

Boer, G. J. (2000), A study of atmosphere-ocean predictability on long time scales, *Climate Dynam.*, *16*, 469-477.

Boer, G. J., and S. J. Lambert (2008), Multi-model decadal potential predictability of precipitation and temperature, *Geophys. Res. Lett.*, *35*, L05706; 05701-05706, doi:doi:10.1029/2008GL033234.

Castro, S. L., G. A. Wick, and W. J. Emery (2012), Evaluation of the relative performance of sea surface temperature measurements from different types of drifting and moored buoys using satellite-derived reference products, *J. Geophys. Res.: Oceans*, *117*(C2), C02029, doi:10.1029/2011jc007472.

Collins, M. (2002), Climate predictability on interannual to decadal time scales: the initial value problem, *Clim. Dynam.*, *19*, 671-692, doi:DOI: 10.1007/s00382-0254-8.

Emery, W. J., D. J. Baldwin, P. Schlüssel, and R. W. Reynolds (2001), Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements, *J. Geophys. Res.*, *106*(C2), 2387-2405, doi:10.1029/2000jc000246.

Frank, P. (2010), Uncertainty in the Global Average Surface Air Temperature Index: A Representative Lower Limit, *Energy & Environment*, *21*(8), 969-989.

Frank, P. (2011), Imposed and Neglected Uncertainty in the Global Average Surface Air Temperature Index, *Energy & Environment*, *22*(4), 407-424; open access: http://multi-science.metapress.com/content/t408x847248t411126/fulltext.pdf (847241 MB).

Hadfield, R. E., N. C. Wells, S. A. Josey, and J. J. M. Hirschi (2007), On the accuracy of North Atlantic temperature and heat storage fields from Argo, *J. Geophys. Res.: Oceans*, *112*(C1), C01009, doi:10.1029/2006jc003825.

Hubbard, K. G., and X. Lin (2002), Realtime data filtering models for air temperature measurements, *Geophys. Res. Lett.*, *29*(10), 1425 1421-1424; doi: 1410.1029/2001GL013191.

Hubbard, K. G., X. Lin, and E. A. Walter-Shea (2001), The Effectiveness of the ASOS, MMTS, Gill, and CRS Air Temperature Radiation Shields, *J. Atmos. Oceanic Technol.*, *18*(6), 851-864, doi:doi:10.1175/1520-0426(2001)018<0851:TEOTAM>2.0.CO;2.

Jiang, J. H., et al. (2012), Evaluation of cloud and water vapor simulations in CMIP5 climate models using NASA "A-Train" satellite observations, *J. Geophys. Res.*, *117*(D14), D14105, doi:10.1029/2011jd017237.

Kawai, Y., H. Kawamura, S. Tanba, K. Ando, K. Yoneyama, and N. Nagahama (2006), Validity of Sea Surace Temperature Observed with the TRITON Buoy under Diurnal Heating Conditions, *J. Oceanogr.*, *62*, 825-838.

Kiehl, J. T. (2007), Twentieth century climate model response and climate sensitivity, *Geophys. Res. Lett.*, *34*(22), L22710,22711-22714;, doi:10.1029/2007gl031383.

Klein, S. A., Y. Zhang, M. D. Zelinka, R. Pincus, J. Boyle, and P. J. Gleckler (2013), Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator, *J. Geophys. Res.: Atmospheres*, *118*(3), 1329-1342, doi:10.1002/jgrd.50141.

Lauer, A., and K. Hamilton (2013), Simulating Clouds with Global Climate Models: A Comparison of CMIP5 Results with CMIP3 and Satellite Data, *Journal of Climate*, *26*(11), 3823-3845, doi:10.1175/jcli-d-12-00451.1.

Levitus, S., J. I. Antonov, T. P. Boyer, R. A. Locarnini, H. E. Garcia, and A. V. Mishonov (2009), Global ocean heat content 1955–2008 in light of recently revealed instrumentation problems, *Geophys. Res. Lett.*, *36*, L07608, doi:doi:10.1029/2008GL037155.

Lin, X., K. G. Hubbard, and C. B. Baker (2005), Surface Air Temperature Records Biased by Snow-Covered Surface, *Int. J. Climatol.*, *25*, 1223-1236; doi: 1210.1002/joc.1184.

Roy, C. J., and W. L. Oberkampf (2011), A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing, *Comput. Methods Appl. Mech. Engrg.*, *200*(25-28), 2131-2144, doi:http://dx.doi.org/10.1016/j.cma.2011.03.016.

Saur, J. F. T. (1963), A Study of the Quality of Sea Water Temperatures Reported in Logs of Ships' Weather Observations, *Journal of Applied Meteorology*, *2*(3), 417-425, doi:doi:10.1175/1520-0450(1963)002<0417:ASOTQO>2.0.CO;2.

Vasquez, V. R., and W. B. Whiting (1998), Uncertainty of predicted process performance due to variations in thermodynamics model parameter estimation from different experimental data sets, *Fluid Phase Equilibria*, *142*(1-2), 115-130.

Vasquez, V. R., and W. B. Whiting (2005), Accounting for Both Random Errors and Systematic Errors in Uncertainty Propagation Analysis of Computer Models Involving Experimental Measurements with Monte Carlo Methods, *Risk Analysis*, *25*(6), 1669-1681, doi:10.1111/j.1539-6924.2005.00704.x.

Wilks, D. S. (Ed.) (1995), *Statistical Methods in the Atmospheric Sciences*, 464 pp., Academic Press.

Williams, K. D., and M. J. Webb (2009), A quantitative performance assessment of cloud regimes in climate models, *Climate Dynamics*, *33*(1), 141-157, doi:10.1007/s00382-008-0443-1.

Xu, F., and A. Ignatov (2013), In situ SST Quality Monitor (iQuam), *J. Atmos. Oceanic Technol.*, *31*(1), 164-180, doi:10.1175/jtech-d-13-00121.1.