IPCC Review Comments: Second Order Draft (SOD) of the Working Group I Contribution
Submitted November 29, 2012 by Ross McKitrick

Chapter 1- Introduction

- [Page 9] This paragraph refers to Figure 1.4, which is described as a comparison of observed globally-averaged temperature anomalies from 1990 to 2011 to the predictions of climate models for the same interval as presented in the sequence of IPCC reports. The Figure is nicely constructed and very informative. But the text in this paragraph does not describe what the Figure shows. For instance it refers to a smoothed black line as showing the observations, but there is no black line in the Figure, instead there are points and confidence interval whiskers. More importantly, the text claims that the observations are inside the uncertainty ranges of past assessments and generally lie in the middle of the model projections from past IPCC reports. This is not the case in the Figure I have. The Figure in my copy of the chapter shows something completely different. Leaving aside the gray shading for a moment (on which see the next comment), the path traced by the observations (black dots) and their uncertainty whiskers, representing the span of observations, dips at the time of the 1992 Pinatubo explosion, jumps at the 1998 El Nino, and then drifts horizontally, with no upward trend over the past decade, while the model forecasts trend steadily upwards, so that by 2012 the envelope of previous IPCC temperature projections lies completely above the observed global average. Are we reviewing the same Figure as the one the authors were writing about?

- [Page 39] The gray shading is defined in the Figure caption as the uncertainty band for the HadCRUT4 temperature series. However, the caption also describes each year's uncertainty band for both the observational series as the black whisker, which is said to span the 90% uncertainty range. The black whisker is typically about 0.15C in length (by eyeball). The gray shading, by contrast, spans a much larger range, namely 0.8-1.2 C after 1992. And the shaded area is trending upward over the past decade while the actual temperature data trends slightly downward. How can a 90% uncertainty interval move upward while its mean value stays constant or declines? Also, the gray shading spans such a wide area as to imply that every year's global average temperature since 1990 lies within the 90% uncertainty band around the current global average, implying that there has been no statistically significant change in the mean temperature for over 20 years. Worse, the shading continues out to 2015, 3 years past the end of the observations on which it is supposedly based! The caption must be wrong. It looks like the gray shading is the result of applying an uncertainty estimator to the model projections, not the observations. That would make sense of the fact that the models roughly track the middle of the gray shading, but all that would show is that the models roughly match the models. The gray-shaded portion is not well defined, it distracts from the Figure, and seems to have been erroneously constructed. I would suggest it be removed altogether.

- [Page 40] Regarding Figure 1.5 and the accompanying text, which apparently focuses only on the AR4 model runs, the same problem arises: the text does not match the diagram. Over the interval shown the observed temperatures are more or less constant and the model projections go upward at an accelerating rate, so that by the end, the observed data are at or below the low end of the range of warming projections. The gray shading is again a mystery. The caption suggests the same definition as in Figure 1.4, but the Figure shows the shading

continues into the future, suggesting it is an uncertainty band around model runs, not observations. So it provides no support for the claim that the observations match the models.

- [Page 2] Assuming that the discrepancy between the text and the figures is resolved by making the text match what is shown in the figures, the summary sentence will also need to be revised. It should say something like "Although CO2 concentrations in the atmosphere have risen in line with earlier projections, globally-averaged temperature observations have risen less than projected and are currently at or below the low end of the range in past IPCC assessments." Or maybe, "As shown in Figures 1.4 and 1.5, since the end of the 1992 Pinatubo volcano, models have predicted a steady upward trend in global average temperatures, but the observed series have been comparatively trendless, and thus the range of model warming predictions since the early 1990s can be seen to have been biased towards more warming than was subsequently observed."

Chapter 2- Observations: Atmosphere and Surface

- [Page 7] The statistical problem of handling long term persistence when computing the significance of temperature trends is of enormous importance for the detection and measurement of climate change, so the continued refusal to deal with this topic in IPCC reports is very unfortunate, especially in light of the massive literature now available on the subject. This paragraph sets the topic aside with the insinuation that researchers can get pretty much get any result they want by choosing their statistical model. Of course the first thing that comes to the reader's mind is that this, presumably, applies to your calculations too. Also, your claim is not demonstrated in the 3 papers you cite. The Mann (2011) paper argues that persistence-like behaviour can be generated in a relatively simple model driven by stationary noise, and then recommends stationary AR representations for statistical modeling on the grounds of Occam's razor. This, however, does not address the points made in the other 2 papers (Cohn and Lins 2005 and Mills 2010), which estimate models that can nest simpler AR structures as restricted cases and show they can be rejected. Occam's razor does not justify using a simple model when it can be explicitly rejected against a more general form. Mills 2010 does state that where rival models cannot be nested a choice must be made, but that is not the point at issue here since he is only comparing more complex models than the ones you use, each of which nest your AR1 model and reject it. Since the use of an oversimplified model, such as AR1, exaggerates the significance of the trends, your continued reliance on it even after it has been long superseded in the expert literature risks providing misleading inferences. You need to present a valid statistical modeling framework, defend it based on a reasonable range of time series modeling tests, and then explain why, if your findings differ from those of others, yours should take priority.

- [Page 181] This Figure's title is misleading. It claims the listed events are "happening everywhere" and the list includes Tropical cyclone frequency and intensity, Heavy precipitation, Cold nights, Heatwaves, Warm days/nights and droughts. But then a more careful study of the graph shows that 4 of the 7 items are stated with medium or low confidence, 2 of the 7 items are going down rather than up, and one is going up and down at the same time. In other words, its detailed content runs more or less opposite to the impression created by the headline. The heading should be changed to something neutral like "Geographic spread of selected changes in weather and climate indicators".

- [Page 4] There are very few specific, numerical estimates in the Chapter summaries and the SPM itself. So those that do get included really should be the most solid and relevant ones.

The claim that unresolved non-climatic biases in land surface temperature data amounts to no more than 10% of the global trend is a guess. Expressing it numerically gives it the appearance of having a scientific foundation, but it is conjured from thin air in the text, it is not an estimate derived from actual data analysis. It is a guess coming at the end of a review of the underlying literature that, among other things, concedes that a previous IPCC report had no supporting evidence for its dismissive assertions on the topic. What you now say in the text is that the issue is disputed and the conflicting lines of evidence have not yet been resolved. That is what the summary should say. You should not just go around making up numbers.

- [Page 20] You say that the confidence interval for your temperature trends is solely that arising from uncertainty in the trend parameter, but you also say that you use the Santer et al. (2008) AR1 method. In light of ample literature in long term persistence that shows that AR1 trend models likely underestimate the uncertainty of trend parameters, you should include a caution to readers to this effect. Or better yet you should use a more accurate method.

- [Page 20] Since I already criticized the use of the AR1 model for estimating trend significance in my comments on the First Order Draft (apparently to no avail), I will just restate the main points here. Obviously the AR1 method presented here is invalid and the treatment needs to be brought up to date. Surely you can't propose to use the "effective dof" method employed in Santer et al. (2008). It is an approximation first developed in the 1930s before computers were available, it is known to be inaccurate for higher-order AR processes and it is incorrect for the purpose of comparing trends among data sets even in the AR1 case. For a recent survey of modern univariate trend modeling issue see Mills (2009) and for a good treatment of the comparison of trends across multiple autocorrelated data sets see Vogelsang, Timothy and Philip Hans Franses (2005) Testing for Common Deterministic Trend Slopes. Journal of Econometrics 126 (2005) 1—24. A standard, but somewhat dated, method would be ARMA(p,q) errors, which are available in any stats software and would at least bring the IPCC up to a 1970s level of statistical sophistication. And you could easily use Newey-West standard errors which are robust to any form of autocorrelation, and which are also available in most modern stats packages now.

- [Page 120] Considering how tight the space restrictions are, why does the IPCC present 5 pages of elementary derivations of trend slope estimators that can be found in any introductory regression textbook? And why go through such lengthy explanations of methods that are 50 or more years behind the state of the art anyway? You are calling a great deal of attention to one of the weakest aspects of your work. All you need to say is, for example, that you use least squares trend estimates with AR1 errors, and then you can cite an intro regression text if somebody wants to see the derivation. Readers who don't know any stats won't read this section, and those who do know their stats won't think much of it.

- [Page 30] Here you say that estimates of large-scale temperature trends have either avoided urban sites or applied corrections based on urban-rural comparisons. But as in my review comments for the FOD, I question how the first claim can be made since CRU, GISS and NOAA all rely on the GHCN archive, and over 60% of the recent Southern Hemisphere land surface data come from urban airports. So they are not "avoiding" urban sites, and it is even questionable whether there are many countries that have enough rural data to support the estimation of corrections. If you have any studies that specifically "avoid urban sites" then put the citation of that study at that specific point in the sentence, rather than nesting the

citations all together at the end of the sentence, where readers can't tell which, if any, support the former claim.

- [Page 30] This paragraph refers to the claim in the AR4 that the observed correlation between the spatial pattern of warming and the spatial pattern of socioeconomic change becomes statistically insignificant after controlling for the effects of atmospheric circulation changes. That claim was the basis on which the AR4 set aside concerns of non-climatic bias in the surface temperature data, which in turn was an essential assumption for many other conclusions in the AR4. You now acknowledge that this claim was made without any supporting evidence, which is quite an admission. You also need to mention that the only paper to investigate the matter concluded it was not only unsupported but false. You cannot create the insinuation that supporting evidence might have existed but was not provided, especially since in the references at the end of the chapter you list my 2010 paper "Atmospheric Circulations Do Not Explain the Temperature-Industrialization Correlation", but do not mention it in the context of this discussion. So a sentence should be added saying something like "The claim was tested in McKitrick (2010) and shown to be likely untrue." You should also list all the follow-on claims in the AR4 that depended on the assertion that the land record was unbiased.

- [Page 30] Your summary of the state of play regarding analysis of evidence of contamination of the surface temperature record is reasonable, with one exception. You say a "hypothesized residual warming artefact..." is inconsistent with models. It's not "hypothesized", it's "observed", as your own text earlier noted. So if it's inconsistent with models, that may indicate a problem with the models. In any case, the wording here is much better than the unfortunate tone of earlier assessments. But you need to follow your own argument to its logical conclusion. You have now clearly indicated that several different methods have been applied, they have yielded substantially different conclusions about the data quality issue, and because they are incommensurable in their approaches it is not currently possible to adjudicate the debate. That is the plain meaning of these paragraphs, and I think it does justice to the literature. But, having admitted the matter can't be decided, you go ahead and do so anyway by making up an arbitrary number, concluding it is "likely" that residual biases are no more than 10% of the underlying trend globally and 25% regionally. Where do these numbers come from? Pulling a number out of thin air doesn't turn a guess into a quantitative science. The sentence would be less inconsistent with the 2 paragraphs that precede it if you say "Notwithstanding the foregoing, we will assume that the magnitude of bias is small enough to be ignored, and we hope that turns out to be the case."

- [Page 158] I would like to reiterate something I said in the previous review round. The Figure showing land temperature changes has a vertical scale from -1.2 to +1.0 C. The oceanic data over the same interval only spans -0.9 to +0.4 C. By stretching the scale it makes the two look like they have the same trend, whereas the Sea Surface Temperature trend is much smaller than that for the land temperatures. This should be illustrated clearly for the reader by keeping the vertical scale consistent, rather than making the two look artificially alike.

- [Page 171] There is an error in the scale for the top left panel. The lower limit should be -1.0, not -0.1

- [Page 171] The figure is supposed to contain global indicators. The one in the 4th row, second column shows declining polar sea ice, but only includes Arctic sea ice extent. It should also include Antarctic sea ice data, in which case I expect the overall trend will be flat.

- [Page 171] The Figure is created to emphasize the similarity of evidence of warming across the different data sets. But some of that similarity is obtained by the fact that the scales are different. For instance, the sea surface and marine air temperatures go from -0.6C to +0.4C, which is only half the range of the land surface temperatures, which go from -1.0C to +1.0C. Since they are measuring the same thing, in principle, they should be shown on the same vertical axes.

- [Page 39] The wording here is a bit confusing. When you say it is "virtually certain" the troposphere has warmed globally, but there is only "medium to low confidence" about the rate and vertical structure, the obvious question is, if you have such low confidence about the rate, how are you "virtually certain" it is not zero? I am not arguing that it is zero, what I am saying is that your language, being far removed from conventional uncertainty terminology, is creating paradoxes. You are saying something like, "we are virtually certain the troposphere is warming at a rate we are very uncertain about." Are you trying to say that the confidence interval is wide, but that it does not include zero? Couldn't you just say it that way? I seem to recall the IAC raised this point about contradictory layered statements where you are virtually certain about something for which you have low confidence in the evidence, or vice versa.

- [Page 39] Further on this same point, the case of the tropics matters quite a bit to readers and comes up in Chapter 9 and ought to be discussed more clearly. When you say "confidence is low" in the tropical upper troposphere, it is not clear by that point what you are referring to. Confidence in the data quality, or in the existence of a trend? Given the first, the second follows pretty directly. In MMH2010 (ref. next cell; note it is the Correction not the original paper) the trend in the tropical lower troposphere is significant or marginally significant over 1979-2009 in all 4 data sets, but in the mid-troposphere it is insignificant in 3 of the 4 data sets. So there is a fair degree of consistency in those results.

- [Page 39] R. McKitrick, Stephen McIntyre, Chad Herman, Corrigendum, Atmospheric Science Letters, 2011, 12, 4

Chapter 9- Evaluation of Climate Models

- [Page 3] There's a discrepancy between Figure 1.4 and this paragraph. You say there is "very high confidence" that models simulate realistically the global surface temperature trend especially over the last 50 years. But Figure 1.4 shows the models predict a uniform upward trend in the global average after 1990 (except the dip at the 1992 Pinatubo volcano, which was programmed in later), while the observed global average bounces around then runs flat for a long time, dropping out by 2010 below the low end of the model range. The discrepancy is obvious in that diagram. Your claim to have "very high confidence" in the model forecasts therefore needs explanation. You cite Figure 9.8 as support for the claim of a match with observations, but that Figure doesn't support it at all, in fact it shows that during the past few decades the match has gotten worse. It shows that during the pre-1992 interval, when you got to peek at the answer, the models matched (or were tuned to match)

the observed trends. But after 1990 when the models could be said to have begun generating forecasts, the model temperatures clearly trend upwards while the observations do not, and the discrepancy opens up so much that by 2011 the observed mean temperature is at or below the entire range of model projections. This appears to be the longest interval in which the model mean and the observed mean do not cross, and the discrepancy is widening over time. You can't ask readers to accept at face value a claim that the models are doing better and better when the discrepancy is getting bigger and bigger. The language of "very high confidence" in this case appears out of place.

- [Page 7] I have difficulty with your statement that the models being evaluated are "based on" fundamental laws of nature. You go on to say that many of the most important processes (clouds, turbulence, biogeochemical processes, aerosols, precipitation, the carbon cycle, etc.) cannot be represented in terms of their fundamental physical laws so they must be approximated through empirical parameterizations. So I suppose the models are "based on" fundamental laws of nature the way this or that movie is "based on" a true story. It would be more accurate, I think, to say that the models are mainly built up using approximations to fundamental processes that are put together in a way that try to avoid, as much as possible, violations of fundamental laws of nature, especially the laws of conservation.

- [Page 17] This is a very interesting section. If I understand the point correctly, it is saying that the atmospheric component of a climate model is essentially a weather prediction model, and could be used as such if the correct initialization data were available. And, as a consequence, it has been shown that systematic errors in model behaviour develop quickly in the forecast interval. Now there has been a lot of discussion about the failure of observed temperatures to go up over the past 10-12 years, and longer in the tropical troposphere. The response of modelers has been to say that these intervals are too short to decide if there are any systematic errors in models. But here you seem to be saying that with regard to some key model processes, if they are going to go wrong, it will show up quickly. In other words, a discrepancy over a relatively short interval of time would be sufficient to indicate an error. So it would be helpful if we were given some kind of guidance as to what is the time scale necessary for testing the major hypotheses embedded in climate models. For example, how many years of observations on the tropical troposphere are necessary to test whether it is represented incorrectly in models?

- [Page 1] The chapter is called "Evaluation of Climate Models". But the methodology of evaluation seems focused on looking for signs of agreement between models and data. There is no discussion of hypothesis testing. Models embed hypotheses that may, in principle, be wrong. While there are discussions of errors and deficiencies in small-scale local processes, there seems to be a working assumption that anything on the large scale is basically correct in the models, and model evaluation is only for the purpose of looking for signs of fidelity between models and data sets. To the extent discrepancies are observed, no matter how numerous, this does not seem to affect the prior judgment that the models are basically accurate. So, for the purpose of model evaluation, what would constitute a testable hypothesis? If you are going to recommend we have high confidence in the models as forecast tools, presumably they must be falsifiable.

- [Page 26] Regarding the stated range of tropical lower- and mid-troposphere trend estimates from McKitrick et al. (2010), they were updated in McKitrick et al. (2011), where the range is LT: 0.08 - 0.15 C/decade, and MT: 0.03 - 0.11 C/decade.

- [Page 26] Reference: R. McKitrick, Stephen McIntyre, Chad Herman, Corrigendum, Atmospheric Science Letters, 2011, 12, 4

- [Page 26] The point made in the last line of this paragraph is extremely important. As you indicate, the evidence for warming, or lack thereof, in the tropical troposphere has been subject to intense controversy and seems to have considerable importance for assessing climate models. The models all indicate that there should be amplified warming in the tropical troposphere, and it is a vast part of the atmosphere with apparently great importance for understanding planetary-scale climatic processes. So it seems inadequate to observe that some data sets show no significant warming in that region over the 1979-2009 interval and then move onto another topic. Although the next paragraph goes on to compare observed trends to the model projections, shouldn't there first be some discussion of the importance of this region? Wouldn't the absence of any significant warming in this region indicate something about, say, water vapour feedbacks? Recall that the context of this issue is your declaration of "very high confidence" in the models, on the basis of which many of the major conclusions of the IPCC report will be based. If there was a significant warming trend in the tropical troposphere you would no doubt highlight it as a reason for your high confidence in the models. Since you maintain that you have very high confidence in the models it must mean you can explain the lack of warming. This would be a suitable place to do so.

- [Page 27] This statement seems to insinuate that the studies finding a statistically significant discrepancy between models and observations in the tropical troposphere did so by averaging across the observational series. But McKitrick et al. (2010; 2011-ref. in row 26) reported results for individual series as well as for multi-series averages. The discrepancy between models and observations is statistically significant either way: in the case of series averages and for every individual series as well. If you are going to mention the distinction then you need to mention these findings as well.

- [Page 27] The idea that observational trends should be compared to the extrema of model trends, rather than to the confidence interval around the mean of model trends, is statistically and methodologically incoherent. It is noteworthy that you have no supporting citations for this position. It amounts to a recommendation to engage in cherry-picking, and it is contradicted by your own methodologies elsewhere. You have already stated (p. 9_8) that a single model run can follow any one of many pathways. So to characterize the behaviour of a model you usually use ensemble means, presumably to draw out the underlying common aspects of the model runs in a forcing scenario. And you claim that the models are based on fundamental physical laws, implying that there is an underlying core theory common to the models. Presumably that core theory is revealed in the average behaviour across ensemble members. Yet here you say something different: that the proper way to compare models and observations is not to use the means but to use the endpoints of the full spread of the model runs. This is a bit too convenient: you can make that spread as wide as you like simply by adding more and more runs. Given enough runs, even from biased and incorrect models, eventually one will have a trend that coincides with the observed data. This proves nothing in a set-up where you can generate infinitely many model runs. It is not evidence in support of the "fundamental laws of nature" upon which the models are based, nor does it validate their parameterization and tuning, nor does it support anything to do with the models as a genre. All it says is that if you roll the dice often enough, eventually you get snake eyes. To say something about models as a group you have

to test their average/common trend against the observational counterpart, which is precisely what McKitrick et al. (2010) does. Your argument, in effect, tries to have things both ways. You want to claim very high confidence in "the models" as a unified methodological entity or genre, but then you propose testing them as independent, atomistic single runs. Leaving aside the problem of cherry-picking, even if you got a perfect match between the observed trend and that from model run #1008, it tells you nothing whatsover about the model as a scientific tool, because you are not testing the model, you are just testing a list of numbers that came out of it. If you want to draw a conclusion about the model, you have to treat it as a data generating process and test it as such, which means taking account of the distribution of what it produces, i.e. the moments. The paper you cite as supporting evidence for your method does not address the issues under discussion. In fact the logic of Bayesian Model Averaging goes completely counter to what you are proposing to do, since it is used to neutralize cherry-picking (or "model selection") bias in situations where researchers can pick from an extremely large number of models.

- [Page 27] This sentence makes it sound like the model-observation discrepancies were only found due to an improper uncertainty metric. No, the uncertainty metric was correct. The discrepancies were found because the models over-predict warming in the tropical troposphere, and robust trend estimators indicate that the difference is statistically highly significant, so that the models on average predict a trend that is significantly higher than any individual observational series or all observational series averaged together. You cite no published papers in support of your claim that a better method would use "the standard deviation or some other appropriate measure of ensemble spread." In fact you can't even say what alternative measure you would prefer! Much less do you cite a paper that argues for it and uses it. So you simply are not in a position to ignore or set aside the findings in the published literature. As stated above, if you are going to appeal to the spread of individual model runs then you have to abandon any claim to have validated the models as a genre, or as a collective embodiment of the core theory of how the climate works. To the extent you want readers to think of climate models as a single collective genre, as represented by a core set of processes "based on fundamental laws of nature" it makes sense to talk about the behaviour of the average of model runs. And, as you point out, on that basis the model-observation discrepancies are very highly significant.

- [Page 27] These paragraphs make clear that there is no apparent explanation for the tendency of models to over-predict warming in the tropical troposphere. So you shrug and say that the cause of the bias remains "elusive". Given that this section refers to the lower- and mid-troposphere in the tropics, which makes up about half the Earth's atmosphere and is the place where all the models predict the most rapid and pronounced GHG-induced warming should be observed, how can you then state at the beginning of the chapter (p. 9-3, lines 48-51) that you have "very high confidence" that the models correctly simulate the atmospheric response to greenhouse gas forcing? The most prominent feature of climate models' response to GHG forcing is rapid, amplified warming in the tropical troposphere. Yet in this section you have conceded that it's not even clear whether there has been any warming there at all, and to the extent there has been it is far less than what the models predicted in response to the observed increase in GHG levels, and the model-observational discrepancy is highly significant, and you have no explanation for it. I can't see how you nevertheless have "very high confidence" in the validity of models' handling of GHG forcing, especially in light of the similar discrepancies at the surface as shown in Figures 1.4 and 9.8, in which models predict a lot of warming that does not show up in the surface temperature

record over the past 2 decades. How can you have "very high confidence" in models that get a key prediction wrong over multiple decades for reasons you are unable to explain?

- [Page 21] Check the headings in this section. I think 9.4.1.1.2 is supposed to be 9.4.1.2; also there are 2 9.4.1.3's.

- [Pages 21-32] This is an important section since it discusses the ability of models to get temperatures correct. Section 9.4.1.1 is on the spatial patterns of the mean state. I would have thought 9.4.1.2, or some other subsequent section, would be on spatial patterns of trends. But I didn't see any such section. There is only a discussion of the large, global-scale trends (9.4.1.3) but not one on the spatial pattern of trends, which seems to me important for justifying the "fingerprint" approach to signal detection. The McKitrick and Tole paper in Climate Dynamics (see ref. in cell 34) tests how well each CMIP3 model reproduces the observed spatial pattern of trends over land in the CRU data set from 1979-2002. We used classical and Bayesian methods, testing the models independently, and in every possible linear combination, as well as against alternatives consisting of a vector of indicators of socioeconomic change and exogenous geographic processes. We found that 20 of the 22 models either have no explanatory power or are anticorrelated with the observed pattern of trends; that the socioeconomic variables have unique explanatory power that is not accounted for or encompassed by any GCM; and that a Bayesian Model Averaging exercize involving some 357 million combinations of explanatory variables shows that only 3 of the 22 GCMs (IAP-China, INM-Russia and NCAR CCSM 3.0) account for all the explanatory power among all the climate models. I'm not sure if these findings are of any interest to the chapter authors, but if so, the article citation is in the next cell.

- [Pages 21-32] McKitrick, Ross R. and Lise Tole (2012) “Evaluating Explanatory Models of the Spatial Pattern of Surface Climate Trends using Model Selection and Bayesian Averaging Methods” Climate Dynamics, 2012, DOI: 10.1007/s00382-012-1418-9

Chapter 10- Detection and Attribution of Climate Change: from Global to Regional

- [Page 16] This summary paragraph claims that the spatial patterns of warming from models forced with GHG's and other anthropogenic forcings agrees well with observations. But the underlying text (p. 10.14) provides no statistical tests to support this claim. All it gives is an eyeball comparison of spatial colour maps for the 1901-2010 intervals, and later on the same page notes that the similarity is not as good for the 1979-2010 interval, with evidence of model over-prediction of warming in a number of areas. Statistical comparisons are not provided: readers do not even get a correlation coefficient, let alone a significance test. Nor is any such information given in the one paper cited (Sedlacek and Knutti 2012, which isn't really on point here). For the 1979-2002 interval, extensive statistical tests are provided in McKitrick and Tole (2012, cell 34). Looking at their Table 3, only 2 out of 22 CMIP3 climate models have significant explanatory power for the spatial pattern of warming trends over land, and the rest have no significant explanatory power or are even anticorrelated with the observed trends. This finding emerges whether the models are tested individually, all at once, or in any linear combination. The Sedlacek and Knutti paper is only about oceanic temperatures, not the land record, it shows that the models do a poor job matching observed oceanic changes over the 20th century when relying only on natural forcing, and that if the natural-only runs are scaled to have an overall trend that matches the observations, the models predict a more heterogeneous distribution of trends than was

observed. It's an interesting enough paper, but the argument ultimately depends on the premise that the model is fundamentally correct, so if the natural-only control run doesn't look like the real world, then the natural-only assumption must be wrong. In other words, the paper assumes the spatial validity of the models, so it cannot simultaneously be cited as evidence in support of the same assertion, otherwise you are begging the question. Consequently, if you are going to make a summary statement that the models are able to simulate correctly the spatial pattern of trends (especially after 1979), you need to find some published support, and you also need to address the counter-evidence in McKitrick and Tole 2010.

- [Pages 16-17] Your summary of McKitrick and Tole (2012) is incorrect on two points. First, we didn't just apply BMA, we used two other methods as well, namely encompassing tests and non-nested regressions. Second, we didn't apply the method to both surface and lower tropospheric temperatures, we only looked at the surface patterns, though we used the LT series as a control. You dismiss the findings by stating that in Chapter 2, socioeconomic activity is not assessed to be a major issue for the land data. The use of the passive voice here is noteworthy, since you don't have any published citations to support your position. The Chapter 2 material is more subtle than you make it out to be. They do not overturn any of the evidence of surface data contamination and they note that the disputes are unresolved. Their claim that the problem is relatively small (< 10%) is simply made up at the end of the discussion. So you are compounding the problem by citing their conjecture as evidence for your assumption. I suppose it would complete the circle if the Chapter 2 authors cited your assumption as evidence for their conjecture!

- [Pages 16-17] It is also important to note that the findings of McKitrick and Tole that are relevant for this discussion are not dependent on what you make of the role of the socioeconomic variables in that analysis. The encompassing tests show, with unambiguous clarity, that the explanatory power of the socioeconomic variables is independent of the explanatory power of the climate models. There is no sense in which the explanatory power of the socioeconomic data could be reduced to a spurious effect properly attributable to climatic processes represented in the climate models. They are completely orthogonal to each other: the p-values on this matter are all on the order of 5e-6 and smaller; see Table 4 and cross that escape route off the list. The BMA analysis also allows each group of variables to be considered independently of the others. So for your purposes, the takeaway message is that you can look only at the results pertaining to the GCMs and ignore everything else if you like, and you won't be misinterpreting anything. And the findings in M&T are that only 2 of the 22 GCMs have significant explanatory power for the surface trend pattern, and in the Bayesian sense only 3 have a posterior probability above 20% of belonging in the correct model of the surface temperature trend pattern. So your quick, offhand treatment of the paper, in the context of a chapter that depends heavily on the assumption that the models get the spatial pattern of warming over land correct, doesn't look very sound.

Summary for Policymakers

- [Page 8] It is intriguing to read your categorical dismissal of any role of Galactic Cosmic Rays on climate, expressed with, as you put it, "high confidence". The matter is, of course, an open research question and the work has not yet concluded. It is quite daring of you to declare high confidence in what the results of yet-to-be-completed experiments and analyses will be. But you should consider whether you really want to go so far out on the limb here.

- [Page 9] Here and in Chapter 9 you claim to have "very high confidence" that models provide a realistic response to GHG forcing. But then you go on to say that they don't do very well on precipitation. And in Chapter 9 you admit that the models are significantly off regarding warming in the tropical troposphere, namely that they predict far more warming than has been observed, and you have no explanation why. And Figure 1.4 shows that all the models from past assessments over-predicted warming of surface temperatures over the past 10-20 years. And there is no assessment of the ability of models to get the spatial pattern of trends correct over land, but the published evidence (McKitrick and Tole 2012, cited in Ch 9) shows the models as a whole do very poorly at this. In light of all this, how can you claim to have very high confidence in the validity of the models' representation of the climatic response to GHG's?

- [Page 9] You should insert Figure 1.4 here as part of the summary, and discuss what it shows.

- [Page 20] In this Figure you show indicators of a changing "global" climate, but the ice data is only from the Arctic. You should also show the Antarctic sea ice extent, for the interval for which it is available.