1. The paper's treatment of the significance of the results is still less rigorous than I feel is appropriate for the forecasting topic. For instance, even statements such as (p5 l1) "The HSS above 0 indicates that the forecasts have skill" (positive, negative?) are simplistic. The skill estimate is certainly much stronger (ie significant) if the HSS has been calculated from a sample of 1000 obs-forecast pairs, compared to a sample of 10. I recommend adding to the discussion of the skill score a small general discussion regarding uncertainty (due to sampling error) in the Heidke skill score estimates, but related to the samples sizes used in this paper – ie, 28. It should be fairly straightforward to calculate a confidence interval given this sample size, which can then be referenced in the results discussion (Figures 3-10 excluding 5). For instance, given sampling uncertainty with N=28 the HSS is significantly (positively, $p<0.05$ or $p<0.10$) skillful when it is greater than X (where X is greater than zero).

For the difference between two HSS, the CI may be more difficult to calculate analytically thus the use of the bootstrap is a convenient empirical approach. Yet the authors use of the bootstrap on the sample of 6 mixed scores (3 months + 3 leads) is at least inadequate if not completely incorrect. One strategy would be to generate 1000 trials of two 28-member non-skilled samples (perhaps by shuffling the obs-forecast pairs in time) and calculate score difference thresholds that are exceeded with a desired probability, purely by chance. Another is to bootstrap similarly on the actual 28-member samples to assess the effect of their sample uncertainty on the difference in their scores. Because the HSS is a widely-used metric in weather/climate forecasting, and it is likely a common challenge to assess whether one forecast is better than another, I expect examples can be found in the literature. It is still a striking feature of figures 12-13 that adjacent pixels with different signals (eg -15, +10) or signals of zero are all found significant, when the underlying climate maps (NLDAS and CFSv2) are much smoother. Despite an earlier request for analysis and discussion on this point, this issue remains unexplained.

Finally, the use of the significance calculations in the figures can be improved. Rather than show the significance maps separately (eg in Figure 11), which makes it very difficult to match a pixel's significance and value, the example of skill masking at CPC could be followed – eg, http://www.cpc.ncep.noaa.gov/products/CFSv2/htmls/usPrece1SeaMask.html)

Before this paper can be accepted it will require a more rigorous and thoughtful treatment of uncertainties in the skill score estimates, referencing appropriate literature and clearly describing the approaches used to assign significance.

2. In general, I feel the authors have adequately addressed the reviewers' comments, but I ask that the authors go back through the first round of comments to reconsider and upgrade any perfunctory and limited responses such as the one highlighted below.

Reviewer: Page 2, line 23-25: Please explain how GCM outputs can be used for daily or short-term forecasts seeing as they are uncorrelated to current meteorological conditions.

RESPONSE: Thanks for pointing this out. This is an overstatement. We have changed the sentence to "Coupled Atmosphere-Ocean General Circulation Models (GCMs) are used to make forecasts at multiple timescales."

There are a number of interesting reasons why GCM outputs can and are being used for daily and short-term forecasts when they are initialized for weather and climate prediction, yet these are not discussed in the response or, more appropriately, the paper (as background, perhaps). The reviewer may be confused by the use of GCMs in free-running climate projection mode (where there is no correlation) rather than in operational forecasting mode, and here it would have been appropriate to make the distinction and to describe briefly the sources of GCM predictability at short-to-medium ranges (eg the initializations, inertial dynamics at different scales and in different components, etc.).

3. The writing still requires a careful proof-reading by an accomplished technical English-language writer.