This study conducted an interesting research about three influential approaches in evaluating the climatic effects induced by afforestation over China. So far, no such studies have ever compared the three methods simultaneously and investigated the underlying mechanisms that lead to their discrepancies and more importantly, whether the discrepancies can be mitigated or reconciled. I'm happy to see that the authors filled this knowledge gap and gave us a good reference. As far as I know, in previous studies involving both the actual and potential effects (Li Yan, 2016, JGR-A, Shen Wenjuan, 2019, AFM), the two effects, characterized by LST changes (or cooling) were comparable and consistent in magnitude. As a result, their discrepancies attracted less attention. Fortunately, this research emphasized this point by applying the afforestation experiment over China. Coincidentally, I have a pending research (in prepare for subscription) in support of the result (actual effect is largely less than potential effect) in this study.

Overall, I appreciate the authors' efforts to put this question forward and gave a good demonstration. Yet specifically, I have some comments or questions as follows:

(1) The distribution of sample grids about the actual and potential effect were not shown. Maybe you can display them in Supplemental Materials, like Peng Shushi et al., 2014, PNAS did.

(2) Line 313: Please explain why GlobeLand30 is not suitable for detecting forest change, instead of just citing Zeng et al., 2021.

(3) When computing the mixed and full potential effects, what threshold did the authors use to define a 1-km pixel as afforested pixel using the GLC data? In addition, the method to process land cover data (Globeland30) seems to be ambiguous, since Line 189 described using the majority method to aggregate 30 m to 1km, but Line 309-310 mentioned "vegetation type with area fraction > 50% for every 1km * 1km window". In my opinion, majority does not equal > 50%. For instance, one land cover type (i.e., cropland) accounts for 30% can also be designated as the dominated type as long as 30% is the largest area fraction.
(4) Line 311. What dataset did forest and openland stem from? Based on the early description, forest was only form GLC data and openland only from Globeland30. Please give a clear declaration here. Once more, it’s important to clearly elucidate the criterion to define the afforested 1-km pixel when aggregating 30-m pixels. If the authors used 50% as the threshold, then the bars below 50% in Figure 6 seem to be unreasonable because pixels with afforestation fraction below 50% was not afforestation anymore. But if using a lower threshold, would the 1-km pixel stay as an afforestation pixel? Please, give an explicit and consistent explanation.

(5) When collecting the sample pixels, did the authors consider the impact of water pixels? As far as I know, the common method is to abandon the grids in which water pixels account for more than a fraction (5% or 10% or 15%...).

(6) Section 2.4, I wonder about the significance and necessity of using Bonferroni correction in this study. Many audience including me seem not to be familiar with this operation. The authors may give a more detailed explanation.

(7) Figure 6. When the fraction of afforestation reached (50, 60], why the mixed potential effect exceeded the full potential effect. It seems strange and no explanation about this phenomena was seen. In addition, significant linear trend can be found for actual effect (as displayed in Figure 5), but it seems that this significant trend was not found in mixed potential especiall the full potential effect. May the authors give an explanation about this?

(8) The reconciliation was reached when increasing the fraction to 100% for the actual effect. But why the fraction increase (through linear extrapolation) was only implemented for actual effect rather than both actual and mixed potential effect. It seems unfair because the author compared the 100% fraction-based actual effect with not 100% based (mixed) potential effect.

(9) What is the difference between Figure 8 and Figure A6 ? Mean values of all grids for Figure 8 and gross values of all grids for Figure A6? Do the cumulative biophysical changes only refer to delta_LE? Because the numbers in Line 586-587 corresponded to delta_LE in Figure. A6.

(10) Uncertainty about the Global Forest Cover dataset should be discussed. References can be found in recent papers published by Dr. Zeng henzhong.

(11) The reasons leading to the discrepancies between actual and potential effects were not considered and discussed thoroughly. 1) Actual effect was calculated using the LST data from two years (target and reference year), but the potential effect used the LST from the same year (2012 in this study). 2) When computing the actual effect, the control pixels were constant or stable unchanged forests, however, as for potential effect, the reference pixels were cropland or grassland pixels. 3) Even though the author adopted the same sample pixels (same locations) for the three approaches, the inherent afforestation fraction was not consistent because different criterions were adopted. Please give a detailed explanation and discussion about the above aspects.