Review Comments

Wang et al. developed a gap-filled ET product and compare with reference data products. The authors conducted extensive analysis and the manuscript appears strong. However, I have some major concerns, including: 1) the method used here does not show universally good performance for many sites, as a large number of sites (about 75 out of 339) do not demonstrate good performance, therefore this inconsistency limits the broader applicability of the gap-filled data; 2) the comparison between site ET (either gap-filled and observed) with ET products is problematic due to scale mismatch.

Major Comments

1. L118. "Subsequently, reanalysis products are utilized to fill gaps in the meteorological data from these networks." How are the reanalysis products specifically used to fill the gaps for the meteorological data? It's common for differences to exist between reanalysis data and local site meteorological data. If you cannot ensure the accuracy of the gap-filling of meteorological data, it may not be reliable to use this data to gap-fill ET, as it could introduce biases due to meteorological data inaccuracies.

   Additionally, the gap percentage for different site observations varies (e.g., L185~189). When you gap-fill the meteorological data, on some days you may only need to fill certain variables, while others may have observations available, leading to inconsistency.

   I don't believe directly using these reanalysis data without reasonable adjustment (e.g., downscaling the data to site level) is appropriate, even though you show the comparison in Figure 4. This is a major concern for your method, and it's crucial to address it thoughtfully.

2. Section 2.2.2, you used different variables from different sources of reanalysis datasets. How do you ensure the consistency between these variables? Please elaborate on any methods or checks used to maintain data consistency across different sources.
3. How did the authors consider the spatial scale mismatch between: 1) the reanalysis datasets of ERA5-Land vs GLDAS vs MERRA-2; 2) the ET products from different sources with different spatial resolutions; 3) these datasets vs. the ET observations (which footprint

should be less than the previous datasets)? Consequently, it may not be appropriate to conduct the comparison analysis between gap-filled ET vs ET products, e.g., as shown in Figures 7 and related analysis. Please discuss how you addressed or accounted for these scale discrepancies in your analysis.

4. Figure 5. This figure appears unusual. Overall, the R2 is very low, but only within a certain range does the R2 show good performance. Does this imply that your method is only applicable under specific conditions? If so, please explain these conditions and why the method might perform better under them. Figure 6. Please provide some additional analysis: show a performance summary table for each site, including relevant metadata that might explain performance variations, can put them in the supplementary.
5. Section 4.3 "4.3 Uncertain performance of gap-filled ET". I don't quite understand the title. In this section, you showed 22 sites with poor performance for the gap-filled ET. It seems you are discussing the reasons that lead to the bad performance, rather than the sources of uncertainty. Consider revising the title to better reflect the content of this section.
6. L328. "This decline is likely due to frequent fire events at these sites, as noted by Yang et al. (2023)". Given that you're showing examples of two sites (US-xSJ and ES-LMa), it should be feasible to use existing data (e.g., remote sensing related fire data) to demonstrate that fires did occur at these two sites and induced the LAI and ET decrease. This would strengthen your argument considerably.

Minor Comments

1. Section 2.2.2. Need to add references for the reanalysis datasets used. This will help readers locate more detailed information about these datasets if needed.
2. L64. In this paragraph, the authors introduce many gap-fill methods. You mentioned that "Although existing methods offer viable solutions for filling ET gaps, most lack a robust physical foundation and largely rely on the selection of specific inputs or possess complex model structures. Furthermore, the required input data for these methods are often difficult to obtain, which compromises their applicability and reduces their spatiotemporal scalability." Could you provide a table in the supplementary material to list the comparison between these methods vs the full-factorial method, including the description, equation,

limitations, inputs, and other relevant aspects? This would make it easier for readers to better understand the advantages of your approach in the context of existing methods.