In this paper Lin et al. present fusing satellite embeddings to improve streamflow reconstruction across river networks. The topic is timely and would be a valuable contribution to HESS. The use of satellite foundation-model embeddings as spatial context for streamflow reconstruction is novel, and the controlled comparison between Base, Base+E, Base+ATLAS, and Base+ATLAS+E is a strength of the study.

The results are promising especially the improvements in KGE and NSE with the satellite embeddings included. However, the current evidence supports a more limited conclusion than the manuscript currently claims. In its current form, the study should be framed as a useful proof of concept for one basin and one year, rather than as a general demonstration of transferable streamflow reconstruction across river networks. Broader claims about generalization and scalability would require additional validation or substantial qualification, particularly with respect to the major methodological concerns outlined below.

Recommendation: Major revision

Major concerns

1. **Single-year, single-basin evaluation limits the conclusions**
   The entire study is based on 2017 data from 41 stations in the Yangtze River Basin. Because the analysis is limited to a single year, method's robustness across dry years, wet years, average years or different hydroclimate regimes cannot be concluded. In addition, all gauges are located in the same large basin, so they may share meteorological conditions, seasonal patterns, and spatially correlated model errors. Although authors acknowledged this issue in the discussion section, this limitation should be emphasized more strongly in the abstract, conclusions, and interpretation of the results, because a single-year experiment cannot demonstrate robustness across different hydroclimatic years.
   Recommendation:
   Authors should either add stricter validation experiment or qualify the conclusions. Additional test would be including such as leave-one-tributary-out validation, spatial block cross-validation, excluding hydrologically connected upstream/downstream gauges, multi-year evaluation or cross-basin testing. If additional data is not available as it is mentioned in the manuscript, the study should be explicitly framed as single-year single-basin proof of concept rather than general streamflow reconstruction across river networks.
2. **Hyperparameter tuning may not be independent of the LOSO test stations**
   It is stated in the manuscript that the hyperparameters are tuned using the station-based grouped cross validation and that performance is then evaluated with the leave-one-out cross-validation. In this regard, it is not clear whether hyperparameter tuning was nested inside each LOSO fold. This is a critical issue if hyperparameters were tuned once with all 41 stations and then LOSO evaluation is performed, the test station in each LOSO fold would have influenced model selection and cause information leakage and result in optimistic performance estimates. For a valid pseudo-ungauged evaluation, the test station must be excluded not only from model fitting but also from hyperparameter selection.
   Recommendation:
   Full tuning procedure should be described. If hyperparameters were tuned separately within each LOSO training set, this should be stated explicitly. If tuning was performed globally using all stations, the authors should rerun the evaluation with nested tuning or

provide a sensitivity analysis showing that global tuning does not materially affect the conclusions.

3. **Inconsistencies for the metric definitions**
   The metric section should be corrected. There is inconsistency between the nRMSE definition in the text and in the equation. In text (line 147) it is described as normalized by the mean observed discharge but the equation (equation 8) normalizes the RMSE by the observed discharge range.
   There is also ambiguity in the pBIAS transformation. It should be clearly defined for better understanding. pBIAS is defined as a percentage, but the transformed score is written as 1 − |pBIAS|. If pBIAS is actually in percentage units, transformation is not meaningful. The plotted values suggest that the authors may have used fractional bias rather than percentage bias.
4. **The RiverATLAS comparison is too narrow**
   The manuscript compares satellite embeddings with only six RiverATLAS variables: forest, cropland, and urban fractions at local and upstream scales. This is useful, but it is not a comprehensive comparison against traditional hydro-environmental attributes. RiverATLAS contains many other relevant descriptors, including topography, soils, geology, climate, drainage area, and anthropogenic indicators. Therefore, the conclusion that satellite embeddings provide information beyond traditional hydro-environmental attributes is too broad.
   Recommendation:
   The authors should either expand the RiverATLAS baseline to include a more representative set of attributes or narrow the conclusion to state that satellite embeddings improve performance beyond selected RiverATLAS land-cover fractions.

Moderate concerns

5. **Effective spatial sample size is small**
   Although the model uses daily data, the satellite embeddings and RiverATLAS attributes are static or annual and repeated for each station-day. In this regard, the independent spatial information is closer to 41 stations rather than 41 x 365 daily samples. This is important because XGBoost can be flexible especially with 64 embedding dimensions and many trees.
   Recommendation:
   The authors should discuss this limitation and provide robustness checks, such as station-level bootstrapping, confidence intervals for median improvements, or sensitivity to simpler XGBoost settings with lower tree depth and stronger regularization.
6. **Log-space residual learning should be discussed more clearly**
   Training the residual in log space (line 84–85) is a good choice that reduces the influence of extreme high flows and emphasizes on stable relative corrections across seasons. However, this formulation has consequences that are never discussed in the manuscript. For hydrological applications in flood management or water resources planning at high flows, this can be a systematic underemphasis of the most consequential flow magnitudes.
   Recommendation:
   Authors should discuss the hydrological implications of log-space residual learning and clarify to which application it is aimed or suitable.

Minor concerns

7. **Clarify the spatial support of satellite-embedding aggregation**
   The manuscript should clearly state whether satellite embeddings are aggregated over local incremental reach subcatchments or over the full upstream contributing area. This is important for large downstream gauges.
8. **Clarify zero-flow handling in log space**
   Even if zero flows are rare in the Yangtze Basin, the method should be defined rigorously how zero or near-zero discharge values are handled in log-residual formulation.
9. **Radar plots in Figure 6**
   The radar plots are compact but make quantitative comparison between configurations difficult. Supplementing with paired station-level delta plots of KGE and NSE changes, with confidence intervals on median improvements, would substantially improve the clarity and scientific rigor of the control experiment presentation.
10. **Editorial corrections**
    The manuscript contains several minor language issues, such as “hyro-environmental,” “adding SE lead to,” and “slightly lower the median.” These should be corrected. Reference should be provided for the statement: RiverATLAS information is “outdated and inconsistent”

Overall assessment

This is an innovative and promising manuscript. The idea of using satellite foundation-model embeddings as spatial context for hydrological residual correction is novel and potentially valuable. The results suggest that satellite embeddings can improve streamflow reconstruction in the Yangtze River Basin, especially for KGE, NSE, bias, and variability.

However, the current manuscript needs major revision before publication. The authors should clarify the tuning procedure, correct the metric definitions, qualify the single-year and single-basin scope, address the limited RiverATLAS comparison, and provide stronger evidence that the reported improvements are robust and transferable. With these revisions, the paper could make a useful contribution to HESS as a proof-of-concept study on satellite-embedding-enhanced streamflow reconstruction.