Reply to Referee #1

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

We sincerely thank the reviewer for the constructive suggestions, and for recognizing the novelty of our study. We agree that the present evidence should support a more carefully framed proof-of-concept study rather than broad claims of general transferability. In the revision, we will qualify the title, abstract, conclusions, and interpretation of the results; clarify the hyperparameter tuning procedure and metric definitions; add station-level robustness analyses; and explicitly discuss the current limits of scalability, and application scope.

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.

Response: We agree that the single-year, single-basin design limits claims about interannual and cross-basin robustness. We will revise the manuscript title and key statements to frame the study as a proof-of-concept in the Yangtze River Basin, and will emphasize this limitation in the abstract, interpretation of the results, discussion, and conclusions. We also respectfully note that the Yangtze River Basin is large and hydrologically diverse, and that the 41 gauges constitute a relatively substantial regional station set (Su et al., 2026). However, because consistent multi-year observations and comparable cross-basin data are not currently available for the full experimental design (Lin et al., 2023), we will tone down generalization claims and clarify that the findings demonstrate regional potential.

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.

Response: Thanks for your attention on the hyperparameter-tuning details. In the original experiments, for each input configuration, we used the first LOSO split as a predefined development split: one station was held out, and the remaining 40 stations were used in station-based K-fold validation to select a single hyperparameter configuration. This configuration was then fixed and applied to all LOSO evaluations. This design was motivated by referencing related blocked out-of-sample evaluation practices (Fang et al., 2025; Büechi et al., 2023) and by our goal of conducting controlled comparisons among input settings under a consistent hyperparameter-selection protocol. Because different input configurations have different feature dimensions and information content, hyperparameters were tuned separately for each input configuration using the same development procedure, search space, and tuning budget. The selected hyperparameter setting was then fixed across all LOSO folds for that configuration. This strategy allowed each input setting to be evaluated with an appropriate model configuration, while avoiding station-specific tuning that could introduce additional variation across held-out stations. It is important to note that our aim was to compare the relative contribution of different input information sources, rather than to optimize performance separately for each individual test station.

However, we also fully acknowledge that this is not equivalent to fully nested LOSO. To address this concern, we will revise the Methods section to describe the above details more explicitly. At

the same time, we will add a nested-tuning sensitivity experiment, where each held-out station is excluded from both model fitting and hyperparameter selection. We plan to supplement such information in the SI for readers to have a more comprehensive understanding of the results derived from different hyperparameter-tuning strategies.

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 $1 - |\text{pBIAS}|$. 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.

Response: We thank the reviewer for pointing this out. The description of nRMSE as being normalized by the mean observed discharge was an inadvertent inconsistency in the text. In the actual calculation, nRMSE was normalized by the observed discharge range, consistent with Eq. (8). We will revise the metric section to ensure that the textual description and equation are fully consistent.

We will also clarify the definition of the transformed bias score. The score was calculated using fractional bias rather than percentage bias (pBIAS), so the expression (1 - |bias|) is meaningful and comparable across stations.

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.

Response: We agree that the current RiverATLAS comparison should not be interpreted as a comprehensive benchmark against all traditional hydro-environmental attributes. The six selected variables were intended to represent land-surface and human-modified land-cover conditions, which are conceptually close to the types of information that satellite embeddings may capture. However, we acknowledge that RiverATLAS includes many other relevant descriptors. Therefore,

we will revise the manuscript to narrow the interpretation and state that satellite embeddings improve performance beyond the selected RiverATLAS land-cover fractions, rather than beyond traditional hydro-environmental attributes in general.

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.

Response: We agree that the effective spatial sample size is much smaller than the number of station-day records. We will explicitly discuss this limitation and avoid implying that the spatial information provides 41 × 365 independent samples.

To assess station-level robustness, we will add a bootstrap analysis based on gauges. For each metric and each model comparison, we will first calculate the station-level performance difference between two configurations. We will then resample the 41 gauges with replacement 10,000 times and recompute the median improvement for each bootstrap sample. The resulting confidence intervals will be reported to show whether the median gains remain robust at the station level. This additional analysis will help clarify the uncertainty associated with the limited number of independent spatial units.

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.

Response: Thanks and we agree that the implications of log-space residual learning should be discussed more clearly. Log-space training emphasizes relative corrections and reduces the dominance of extreme high-flow events during model fitting, which can be useful for stable streamflow reconstruction. However, we also acknowledge that this formulation may

underemphasize the most consequential high-flow magnitudes, and therefore is not specifically designed to optimize flood peaks. We will clarify this point in the methods, results interpretation, and discussion. This revision will help narrow the stated application scope and make the hydrological implications of the log-space residual design more transparent.

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.

Response: Thanks for pointing this out. We will revise the Data and Methods section to explicitly state that the satellite embeddings are aggregated over local subcatchments, rather than over the full upstream contributing area. We will also clarify this when interpreting the results, noting that the current analysis evaluates the value of local spatial context captured by satellite embeddings. Upstream-integrated satellite-embedding aggregation can be an issue for study in future works, but is not currently applied here.

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.

Response: Thank you and we agree that the treatment of zero or near-zero discharge values should be defined explicitly for the log-space residual formulation. We will revise the Methods section to state that a small positive constant is added before taking logarithms, so that zero or near-zero discharge values can be handled consistently. We will also specify the value of this constant and clarify that it is used only as a preprocessing step for the log transformation, without changing the overall residual-learning framework.

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.

Response: We agree that the current radar plots provide a compact summary but are less effective for detailed quantitative comparison among configurations. To improve the clarity and rigor of the control experiment presentation, we will supplement Figure 6 with paired station-level delta plots for metrics. These plots will show the station-level changes for two controlled experiments. We will also report confidence intervals for the median improvements based on the station-level bootstrap analysis described above. These additions will allow readers to better

assess the magnitude, direction, and robustness of the performance changes, while the radar plots will be retained as a concise overview of multi-metric performance.

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”

Response: Thank you and we will carefully revise the manuscript to correct grammatical, wording, and typographical errors. We will also check acronyms, terminology, metric definitions, and figure captions throughout the paper to ensure consistency. Regarding the statement that RiverATLAS information is “outdated and inconsistent,” we will provide appropriate references and clarify the intended meaning to avoid unsupported or overly broad wording.

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.

Response: We thank the reviewer for the positive overall assessment and for recognizing the novelty and potential value of our study. We also appreciate the constructive recommendations for improving the manuscript. In the revised version, we will clarify the tuning procedure, correct the metric definitions, more explicitly frame the study as a single-year, single-basin proof of concept, narrow the interpretation of the RiverATLAS comparison, and add station-level robustness analyses. We will also revise the abstract, introduction, results interpretation, discussion, and conclusions to better reflect the scope and contribution of the study. We believe these revisions will substantially improve the clarity, rigor, and positioning of the manuscript.

References

Su, J., Lin, P., Zheng, K., Lei, X., Yin, Z., Hou, A., and Xie, S.: Observation-based spatiotemporal analysis of the evolving flood regulation capacity for the Yangtze River Basin, Geophys. Res. Lett., 53, e2025GL120255, https://doi.org/10.1029/2025GL120255, 2026.

Lin, J., Bryan, B. A., Zhou, X., Lin, P., Do, H. X., Gao, L., Gu, X., Liu, Z., Wan, L., Tong, S., Huang, J., Wang, Q., Zhang, Y., Gao, H., Yin, J., Chen, Z., Duan, W., Xie, Z., Cui, T., Liu, J., Li, M., Li, X., Xu, Z., Guo, F., Shu, L., Li, B., Zhang, J., Zhang, P., Fan, B., Wang, Y., Zhang, Y., Huang, J., Li, X., Cai, Y., and Yang, Z.: Making China's water data accessible, usable and shareable, Nat. Water, 1, 328–335, https://doi.org/10.1038/s44221-023-00039-y, 2023.

Fang, J., Wu, M., Zhang, Z., and Luo, W.: Leveraging AlphaEarth Foundations embeddings for high-accuracy county-scale corn and soybean yield estimation, TechRxiv [preprint], https://doi.org/10.36227/techrxiv.175825526.60198358/v1, 2025.

Büechi, P. E., Fischer, M., Crocetti, L., Trnka, M., Grlj, A., Zappa, L., and Dorigo, W.: Crop yield anomaly forecasting in the Pannonian basin using gradient boosting and its performance in years of severe drought, Agric. For. Meteorol., 340, 109596, https://doi.org/10.1016/j.agrformet.2023.109596, 2023.