Reply to: Referee #1: Jiang, Shijie

We would like to thank the reviewer again for their valuable feedback and address each point separately in the following.

I went through the revised manuscript and the author's response carefully. Most of the big concerns I raised in round one have been addressed, and I appreciate the effort. A few minor items remain that in my view still need attention before the paper can be accepted.

1. The Introduction refers to "integrated gradients" (l117), while the Methods actually describe input × gradient saliency. Please remove for clarity.

Even though usage of integrated gradients is generally possible with DRRAiNN, we used saliency maps in our approach. We replaced the term and citation accordingly.

2. l376, "no systematic difference" ==¿ But no event-based metric is presented, just by visual judgment is not convincing.

We agree and removed the sentence.

3. You consistently highlight "best three out of five seeds." This reads optimistic. I'd prefer to see averages and spread over all five seeds in text, and then keep the three-seed curves for visualization.

We appreciate this concern. Our seed selection is performed exclusively on validation data (not test data), with test evaluation only after selection to prevent cherry-picking. We believe this approach is methodologically sound because it reflects practical deployment where practitioners select the best validation performance, and we are transparent about evaluating five seeds and selecting three. We acknowledge that averaging all seeds would provide different statistical information. On the other hand, we do explicitly discuss initialization sensitivity as a limitation. Our current approach demonstrates the model's potential when properly initialized rather than averaging across suboptimal initial weight configurations.

Our analyses have shown (and we have confirmed this over the last year over and over again) that suboptimal configurations are rare but unfortunately persist - most likely due to the low target data volume (i.e., actual discharge measurements). Reporting 5 out of 5 distorts the performance due to the rare outliers. This is why we think that it is better to stick to the best 3 out of 5 choice using the validation set (not the test set, which we only use to assess and report the performance - that is, the performance that is reported in the paper). We hope that the reviewer agrees that this is the better choice in the end. Thank you

for your consideration.

4. l239, please specify the value of "predefined threshold"

Thank you for pointing this out, we explicitly added the threshold of 1km.

5. The phrase "fully distributed" in the title/Abstract could confuse readers, since the model outputs only at gauges. Maybe qualify this up front in the Abstract to avoid overselling.

We added to the abstract that DRRAiNN estimates river discharge at gauging stations.

Reply to: Referee #2: Nelemans, Peter

We would like to thank the reviewer for their repeated extensive feedback. All points raised are valid and were addressed, which we think again improved the quality of our manuscript.

Note from the reviewer: The current manuscript is a revised version of original submission by the authors, dated 7th of April, 2025, which I reviewed on 8th of April, 2025. Therefore, I will keep my comments brief, as the authors have already addressed the majority of my earlier concerns. I focus here mainly on a few remaining points.

General comments

The authors present DRRAiNN, a ML rainfall runoff model capable of predicting streamflow at multiple locations. The model architecture is interesting, as it is designed to incorporate certain inductive biases. The model shows good performance, and a notable highlight of the study is the authors' approach to identifying which grid cells influence the simulated discharge at specific locations. Furthermore, the manuscript is well written, with a clear and fluent structure. It is concise, focused, and now addresses several important aspects that were missing in the earlier version.

Thank you!

Specific comments

One of my main issues with the previous version of the manuscript was the authors classification of DRRAiNN as a fully distributed model. I appreciate that the authors now clarify why they consider their model fully distributed. Although I personally still disagree, and would classify the model as semi-distributed, I can accept their reasoning.

My other main issue with the previous version of the manuscript was the lack of reporting on the computational efficiency of DRRAiNN. I appreciate the authors adding this information. Although it is discussed only briefly, computational efficiency is not the focus of the study, and the current reporting on it is in my opinion sufficient.

1. Introduction

The introduction is well written and of high quality. Section 1 "Introduction" and section 2 "Related work" of the previous version of the manuscript have been merged into a single section in the revised version. Taking into account this merger, I appreciate the authors condensing the introduction.

The authors have substantially expanded their discussion of related work, offering a more comprehensive and focused overview of recent machine learning developments in hydrological modelling. I appreciate that rather than merely increasing the number of studies cited, they focus on those most relevant to this study. I also highly appreciate the way the related works are discussed: the authors succeed in both tracing the broader development of ML applications in hydrology over recent years and situating their own contribution within this narrative.

Thank you so much!

Line 58: At small-scale, a lysimeter can be used to directly measure overall evaporation (evapotranspiration).

Thank you, we now state that it is difficult to measure spatially distributed evapotranspiration.

Lines 107-109: I appreciate the authors clarifying why the model classifies as fully differentiable, which was one of my issues with the previous version of the manuscript.

At the end of the introduction, I suggest adding a brief sentence noting that the model is compared to the EFAS model for the Neckar River. This can be very concise and will help set the reader's expectations

We agree and added a corresponding sentence.

2. Methods

This section is well written, and I appreciate the authors' efforts in rearranging the subsections. The overall flow has improved significantly, making the presentation more natural and providing the reader with additional context where needed.

The authors have added a brief introduction to this section, which was missing in the previous version of the manuscript. While the idea of including an

introduction is useful, the current text does not effectively set up the content of Section 2. It reads more like a condensed abstract, lacking a clear overview of the topics covered in this section and referencing material that is actually discussed in the following section, which may confuse the reader.

Lines 121-122, "We present ... distributed manner.": I suggest clarifying here that this section, specifically subsection 2.1, provides a detailed examination of the model's architecture and the rationale behind key design choices, since a general introduction to the model was already given in the previous section. Currently, lines 121-122 simply repeat a condensed version of lines 103-107 and do not clearly indicate what the reader should expect from this section. Additionally, it would be helpful to note that the input and output data of DRRAiNN will be discussed in this section, as these aspects are closely related to the model's architecture.

Lines 122-123, "We evaluate ... design choices.": This is not discussed in this section, but rather in section 3. I suggest to remove this sentence.

Lines 123-124, "We demonstrate ... Awareness System.": The actual demonstration of model performance is part of section 3 "Results". In this section the study area, experimental setup, benchmark model, and evaluation metrics are introduced, and this can be mentioned here in the introduction of the section. They are essentially all components required to assess performance, but the results themselves are presented in the following section.

Line 124-126, "DRRAiNN achieves ... modeled dynamics.": This is a summary of your results and should not be mentioned in the methodology.

Thank you very much for these comments, which thoroughly discuss the introducing paragraph of the method section. We agree that this paragraph feels out of place and revised it accordingly.

Lines 131-132, "... a grid that spans the whole catchment area of the river network.": Some readers might mistakenly assume the static maps and meteorological forcings span only the DEM-delineated catchment area, which is common practice. However, one of the main motivations for developing DRRAiNN is that the effective catchment may extend beyond the DEM-delineated area, and DRRAiNN therefore also takes a larger domain as input. I suggest clarifying that the input data covers a larger domain that includes the DEM- delineated catchment area, but also extends beyond it.

We agree and now explicitly state that the grid spans a domain that is larger than the DEM-delineated area.

Lines 141-142: I appreciate the authors clear description of the nature of the input, internal states, and output of DRRAiNN.

Lines 148-149, "Despite being ... self-organizing nature.": I appreciate this important clarification, as it resolves a point that was unclear to me in the original version of the manuscript.

Figure 1: I appreciate the improvements made to this figure, which now provides a clear and solid understanding of the inner workings of the model, especially in combination with the text. My only remaining concern is the two-sided arrow beneath the box labelled "StationGRU," whose meaning is not yet clear.

Thank you for pointing this out, we clarified the meaning of the arrows in the caption as well as in the main text.

The description of the rainfall–runoff model (subsection 2.1.1) is somewhat difficult to follow, though this is understandable given the model's complexity, and I appreciate the authors' effort to explain it as clearly as possible. In contrast, the description of the discharge model (subsection 2.1.2) is clear, concise, and highly effective

Lines 195-196: I find this very interesting; I was already wondering why one would only aggregate embedded runoff at the stations and not at every upstream river grid cell.

I appreciate the thorough and detailed description of the data used in the study, provided in subsection 2.2. The data preprocessing, as well as the temporal and spatial resolution, are clearly reported, addressing one of my concerns with the previous version of the manuscript.

Line 257: Although I appreciate the authors' description of the temperature data, the extreme values convey limited information about local climate conditions. I suggest replacing these with, for example, the mean summer and winter temperatures.

We agree and replaced the values accordingly.

Figure 3 has improved considerably, and the issues I noted with this figure in the previous version of the manuscript have been adequately addressed.

Line 259: I highly appreciate the authors' description of the discharge characteristics.

Lines 281-284: I appreciate the authors pointing this out, and I agree that it is entirely reasonable to leave the evaluation of DRRAiNN under forecast-based conditions for future work.

Line 292: Changing the batch size during training, to counteract the increased memory use by the simultaneously changing truncation length is a very elegant

solution! I appreciate the authors highlighting this aspect of their work.

Thank you!

Line 293: I appreciate the authors mentioning the computational time for a forward pass, as this was not mentioned in the previous version of the manuscript.

Line 300-304: Reading between the lines, I understand that a CNN trained only on the original data may perform worse when presented with a rotated or reflected version of the data. To address this and improve the CNN's generalization, I understand the authors train it on various symmetries of the data. If this interpretation is correct, I suggest adding a brief sentence explaining the issue of data symmetries in CNNs. Additionally, although it is somewhat mentioned in line 304, it may be helpful to more explicitly state that GNNs do not suffer from this problem.

Thank you for pointing that out. We adjusted the paragraph accordingly, which will likely improve understanding.

Lines 315-326: I appreciate the authors' detailed explanation of the differences between EFAS and DRRAiNN, not only in terms of the models themselves but also their inputs, outputs, resolutions, and use cases. I also value their acknowledgment that the comparison of DRRAiNNperformance with EFAS is not entirely valid. However, as the authors clearly state, this is not their intention; their goal is not to outperform EFAS, but to provide a baseline comparison. In my opinion, their reasoning is entirely fair and justifies the comparison made.

Subsection 2.6: I appreciate the authors providing their reasoning for including each of these metrics, as I previously suggested reducing the number of metrics. With this explanation, I understand their rationale for using four different metrics.

Lines 356-357: The NSE does include the bias, but "in normalized form scaled by the standard deviation" of the target variable (Gupta et al., 2009, see https://doi.org/10.1016/j.jhydrol.2009.08.0 That was one of the main motivations to develop the KGE, which incorporates the correlation, bias, and variability independent from each other.

Absolutely correct, thank you! We fixed this accordingly.

3. Results

The reporting of results is clear and thorough, and the figures are well-designed. Especially Figure 6 and Figure 7 have much improved compared to the previous version of the manuscript. I also appreciate the thoughtful interpretation of

the results and model performance, including the explanation for the varying performance of the DRRAiNN model across different seeds and lead times.

My only significant issue with this section, which I also mentioned in my previous review, is the sometimes seemingly conflicting assessment of the attribution maps. I agree that, to a large extent, the attribution maps from DRRAiNN are expected to overlap with the DEM- delineated catchment areas and could thus be used as an indicator of physical plausability. However, as the authors themselves point out, subsurface flow can, in some cases, transcend DEM-delineated catchment boundaries.

Therefore, I suggest that the authors exercise caution when classifying a DRRAiNN model as better solely because its attribution maps show more overlap with traditional catchment delineations than another DRRAiNN model (lines 461-464). To a certain extent, overlap between the DRRAiNN attribution map and the DEM-derived catchment area is expected and is indeed an indicator of physical plausibility. However, beyond a certain point, increased overlap cannot be assumed to correlate with higher physical plausibility.

In case an attribution map is wildly different from the DEM-delineated catchment, it should indeed be assessed lower than one with more overlap. However, once the differences between two DRRAiNN models are small, caution is needed. As the authors note, DRRAiNN may be capable of detecting unobserved subsurface flow paths, though, as they also emphasize, further research is required to confirm this.

I understand the delicate balance between these potentially contradictory considerations, but I urge the authors to avoid rigorously stating that more overlap automatically indicates a better model, especially since they acknowledge that the DEM-delineated catchment map may not capture all flow paths. When the authors state that one DRRAiNN model instance has more overlap than another and should therefore be regarded as having more physical plausibility (lines 461-464), I suspect they are referring to cases where the lack of overlap from the latter model cannot be explained by subsurface flow alone. If that is the case, I suggest clarifying this for the reader.

If the lower overlap could also be due to undetected subsurface flow paths, I suggest the authors refrain from jumping to conclusions and remain open to the possibility that DRRAiNN may have identified such paths (as, for example, noted in lines 426–427).

Thank you for pointing this out again. We agree that our assessment still was not quite right. We now softened our conclusions and adjusted the paragraph accordingly. We also adjusted the paragraph about the ablation studies accordingly.

Lines 404-405: The quality of the observational data at these locations may influence the performance of both models. The authors are in a better position to assess this. If they consider this a potential reason for the reduced model performance at these locations, I suggest mentioning it here

We agree and added data quality to the list.

Lines 410-411: The NSE does include bias and variability, see Gupta et al. (2009, https://doi.org/10.1016/j.jhydrol.2009.08.003).

Again, thank you. Since a direct comparison of KGE and NSE is non-trivial and tells more about the metrics themselves than the stations, we removed this sentence.

Line 415: I assume the authors intended to indicate that darker areas correspond to regions of higher importance.

Yes, indeed. Thank you!

I appreciate the extensive series of ablations conducted by the authors.

4. Discussion

The discussion is well written, providing a clear summary of the work as well as several interesting suggestions for future research. I appreciate the authors' efforts in condensing the discussion.

Thank you so much!

5. Conclusion

The conclusion is well written and concise, and it clearly highlights the added value of the study.

Thank you!

Technical corrections

Line 39: A period is missing after "biases".

List of relevant changes made in the manuscript

1. Mention that we predict at gauging stations in abstract
2. State that evapotranspiration is difficult to measure (in contrast to impossible)
3. Mention saliency maps instead of integrated maps in introduction, as this is the technique we used
4. Improve transition from introduction to method chapter
5. Explicitly state that our domain in larger than the elevation-delineated catchment area
6. Describe the two arrows between Segment- and StationGRUs
7. Add threshold that was used to remove stations if coordinate correction is too large
8. Add mean temperature of summer vs winter instead of min and max temperatures
9. Briefly explain why CNNs benefit from data augmentation while GNNs don't need it
10. Correctly handle differences in KGE and NSE metrics
11. Explicitly state that lower Wasserstein distance could be due to correctly inferred underground flows
12. Add data quality as possible explanation for performance variance across stations