# Reply to: RC2: 'Comment on egusphere-2024-4119', Peter Nelemans

*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. Overall, the study is well-executed, but at times, the manuscript appears to lack important details, which could pose challenges for readers attempting to build upon this work.*

Thank you very much for your very extensive and detailed comments! We think addressing the raised issues will significantly improve our manuscript. In the following, we will reply each point in more detail.

*Specific comments*

*One of my main issues with the manuscript is its characterization of DRRAiNN as a fully distributed model, even though discharge is estimated only at specific stations. While the input is fully distributed and the rainfall runoff predicts embedded runoff in a distributed manner, no attempt at interpretation of these embeddings was done, so I assume they may not be interpretable. Consequently, if the only physically meaningful output is generated at discrete locations, the model should in my opinion be classified as semi-distributed. I suggest that the authors either provide additional justification for considering DRRAiNN fully distributed or reclassify it as a semi-distributed model.*

Thank you for this suggestion. It seems to us that what "fully distributed" exactly means is a matter of definition. It is certainly true that our model does not produce a runoff estimation in each grid cell. However, the supplementary material shows that our model propagates quantities along the landscape, indicating that the individual grid cells actually contain information about runoff. How this information could be directly translated into runoff is not at all obvious, since we don't have measurements in each grid cell. So, in a sense, we think that our model is already "as distributed as can be," at least when it comes to modeling the real world. We are considering adding the term "river discharge" to the title of our manuscript to emphasize the fact that the estimated quantities are situated within the rivers.

*Another one of my concerns with the manuscript is the absence of any reporting or discussion on the computational efficiency. Although it does not have to*

*be a large part of this study, it remains relevant. This is even more so when considering scaling up DRRAiNN to larger areas, increasing temporal or spatial resolution, or applying the model in flood forecasting scenarios where many weather forecast ensembles must be processed in short time periods.*

This is important, so we will add this information to our manuscript. While the training is of course time consuming (but manageable with only 8 hours on an A100), the time it takes for the model to estimate a day of discharge is in the order of milliseconds.

*1. Introduction*

*The introduction is generally well-written. I suggest that the authors consider emphasizing the "dirty little secret of hydrology": the common assumption that there is no leakage into or out of a simulated catchment through the subsurface. This issue could potentially be addressed by DRRAiNN and represents one of the most interesting and significant contributions of the study.*

Thank you, we will emphasize this aspect more.

*The authors present the model as fully differentiable. Here, "differentiable" does not refer to the ability to perform gradient-based optimization, but rather to a specific approach of combining physics with ML, as defined by Shen et al. (2023, see https://doi.org/10.1038/s43017-023-00450-9). The authors do not elaborate on the distinction between a model that is simply "differentiable" and one that is "fully differentiable", nor do they explain why DRRAiNN qualifies as the latter. While I am not suggesting that it is not, I suggest that the authors provide a clearer explanation of these terms and justify labelling DRRAiNN as fully differentiable, especially given that this feature is highlighted in the manuscript's title.*

With stating that we are presenting a "fully differentiable" model we meant to emphasize "end-to-end" differentiability, meaning that gradients can flow seamlessly through the entire system, enabling optimization of all its parts. We will clarify this in the future version of our manuscript.

*Notwithstanding the two aforementioned (non-binding) suggestions, I find the introduction to be somewhat lengthy. Although there is in my opinion not a specific paragraph that must be removed in its entirety, I do suggest condensing most, if not all, paragraphs to some extent.*

As the other referee and we ourselves agree with this, we will condense the Introduction and will avoid giving too much background.

*Lines 30-31: A lumped model implies spatial averaging at the basin scale but does not necessarily imply temporal averaging. This is also clarified later in the*

*manuscript (lines 116–118)*

Thank you, we removed time from that sentence.

*Lines 31-32, "Therefore, the outline .... be feasible.": The outline of the basin must always be available for PBM-based catchment modelling, regardless whether one employs a lumped, semi-distributed, or fully distributed approach.*

We suggest to differentiate between the approximate and exact outline. Indeed, we also needed know the approximate catchment area of the Neckar as a whole to be able to place our grid. However, with our approach, we do not need to know the exact outline of each sub-basin, as this can be inferred by the model itself. We will adjust our manuscript accordingly.

*Lines 49-50, "...for which no detailed land...": I suggest replacing "no" with "few".*

*Lines 59-60, "[ANNs] high complexity also often leads to neural networks not respecting physical laws despite very good performance": I would argue the complexity of ANNs is not the reason they have difficulty to respect physical laws. Rather, their data-driven approach is.*

*Lines 69-70, "[A] considerable amount of runoff is situated below the ground and therefore not observable": I suggest clarifying a bit: "[A] considerable amount of runoff is often situated below the ground and therefore not directly observable."*

*Line 70, "It is not yet possible to see through the ground": I suggest rephrasing to: "It is not yet possible to directly measure subsurface flow."*

*Line 78, "A combination of the above-mentioned approaches could leverage the advantages of both worlds": I suggest to clarify that you are referencing PBMs and ML here.*

*Line 96, "Another one": I suggest to clarify: "Another inductive bias"*

Thank you for these very detailed suggestions. We incorporated all of them into our manuscript.

*2. Related work*

*This section is well-written but does not sufficiently address the relevant literature. Given that the paper focuses on ML-based hydrological modelling, I suggest expanding this section to include additional studies that explore models with architectures other than LSTM and CNN-LSTM. Without this, readers might mistakenly get the impression that the field has only explored these particular architectures.*

*Specifically, I suggest discussing studies on the use of neural ODEs (e.g., Höge et al., 2022, https://doi.org/10.5194/hess-26-5085-2022), as these models focus strongly on knowledge discovery and system understanding. Furthermore, I suggest including studies that propose physics-informed, fully distributed ML models. E.g.: Sun et al. (2022, https://doi.org/10.5194/hess-26-5163-2022) and Chen et al. (2022, https://doi.org/10.1145/3534678.3539115). Otherwise, this section might unintentionally imply that the pursuit of physics-informed ML models is a recent development.*

Thank you for these suggestions. Unfortunately, we were not fully aware of this additional literature, but think it is important to always consider alternative approaches. We will integrate them accordingly into our manuscript and point out the differences to our approach.

*Line 121, "[F]ully distributed models operate on a grid without any spatial aggregation": In some cases, such as in this study, the forcings are still aggregated, despite the inputs being used in a fully distributed manner.*

We are not certain whether we understand this comment correctly. Does it point to that fact that runoff is collected, and thereby aggregated, at the station locations? An alternative interpretation is the fact that we are coarsening our grid before feeding it into the model. In both cases, we will remove the "spatial aggregation" from this sentence.

*Line 124: I would specifically mention that the model presented by Xiang and Demir (2022) is a GNN, especially given the use of a GNN in this study. Readers might be familiar with GNNs, but not with the work by Xiang and Demir (2022). That being said, I do not think there is a need to go into detail. A simple addition would suffice. E.g.: "The GNN-based model presented by Xiang and Demir (2022) indeed..."*

*Perhaps this is a matter of personal preference, but I suggest placing the authors' names outside the parentheses when referring to them directly. E.g. for line 130: "In contrast, Oddo et al. (2024) used a..."*

Thank you for these suggestions. We incorporated them into our manuscript.

*3. Method*

*I suggest starting this chapter with an introduction to the model and moving what is currently section 3.3 to the beginning. I have several reasons for this. First, the DRRAiNN model is the central focus of the paper, so I would begin the chapter with its introduction. The study site is less relevant to most readers and, in my opinion, would be better placed later in the chapter.*

*Secondly, reordering the sections in this way would improve the overall flow.*

*The current description of the DRRAiNN model is, as it should be, general and independent of the study area or data sources. After introducing the model, the study site and data sources can be presented, as they are more specific. This creates a natural lead into section 3.4, the experimental setup. The chapter would thus progress smoothly from the general workings of the model to the specific details of the experiment, rather than jumping between the two.*

*Lastly, introducing the data after the model would provide better context. For example, in lines 170-171, the data used to construct the adjacency matrix is mentioned. However, this is also the first time the adjacency matrix is introduced, and its function only becomes fully clear in section 3.3.*

Thank you for this nice suggestion! We definitely agree that a reordering would improve the flow and will implement this change in our revision.

*As it stands, the exact role of the different kinds of data in the model is, in my opinion, sometimes difficult to ascertain. For instance, while it can be inferred that solar radiation and precipitation together form the dynamic meteorological forcings F, this is not explicitly stated. Similarly, the static maps S, which, as far as I can tell, only include the DEM, suffer from the same issue.*

We agree and will make better use of the introduced notation to improve intelligibility.

*I also suggest to add a sentence on the study area's climate and some of the characteristics of the river itself (mean, max, min discharge).*

We will add some information about the catchment's characteristics to our manuscript.

*Lines 152-153: Temperature is typically included in hydrological models for two main reasons: to model snow (and glacier) processes, and to model evaporation. Since snow is a rare occurrence in the study area, as far as I can tell (which highlights the importance of including climate details for the catchment), the lack of improvement when including temperature in the input data may be more related to the specific study area than to the model's internal mechanics.*

*Additionally, solar radiation is strongly correlated with temperature, so for evaporation modelling, it might be sufficient to include either solar radiation or temperature, as both could potentially inform the model's prediction of evaporation. While I have no issue with the approach or the sentence itself, I suggest adding some further explanation.*

We assume that our model did not improve when adding temperature is due to the correlation between temperature and radiation, as mentioned. However, we can see that it is not obvious why we chose radiation over temperature

in the first place. This design choice was driven by the intention to model evapotranspiration as well. Nevertheless, we will perform another experiment where we replace radiation with temperature.

*Figure 1: Please include a colour bar for the DEM. For the x- and y-axes, I recommend using latitude and longitude, as the kilometre scale on both axes serves no function given its arbitrary origin, other than acting as a scale bar (which should therefore be included when switching to a lat-long axis).*

*Additionally, the river network is shown in great detail with a consistent width, which makes it difficult to discern which segment a hydrological station belongs to, as the red dots indicating the station locations sometimes overlap multiple streams. To improve clarity, it might help to omit small side streams and/or adjust the width of the river segments according to, for example, the Strahler order.*

*Furthermore, the grey arrows are not included in the legend. While I assume they are meant to indicate the hierarchy between the stations in relation to the flow of the river, I believe they make the figure more confusing rather than clearer. These arrows may be redundant and could be removed if the river network were illustrated more clearly, as I suggested earlier.*

*Lastly, the figure is not referenced in the text. While this might be personal preference, I suggest to include a reference to the figure, for example, in line 141.*

Thank you, these are some very nice suggestions. We will invest more time in improving the figure and implement your recommendations.

*Line 157: I suggest mentioning the temporal resolution as well.*

*Lines 160-161: For additional clarity, I suggest to specifically mention that the mean and standard deviation are computed after having added 1 and taking the logarithm.*

*Line 161: I suggest using "zeros" instead of "0s". Additionally, the fact that 0 is the mean is a result of the standardization process, not the logarithmic transformation. Therefore, I suggest to refer to 0 as the "standardized mean" rather than the "log-standardized mean".*

*Lines 163 and 168: The authors mention using rasterio to transform data to the RADOLAN CRS twice. For conciseness, I suggest stating this once at the end for both datasets.*

*Line 167: I suggest mentioning the spatial and temporal resolution.*

Thank you, we will implement all of these suggestions in our manuscript.

*Lines 175-176: If the discharge measurement station locations are corrected manually, it seems unnecessary to snap them to the river network afterward. I would expect the corrected locations to already be on the river network.*

The station locations were corrected manually by using multiple sources of information, partially from different institutions and maps. Snapping them to the river network actually changed their coordinates, but only to a very small extent.

*Line 188: I suggest removing "future", since historical forcings are used.*

*Line 206: I suggest clarifying that the process described between lines 200-206 is repeated per timestep in an auto-regressive manner.*

*Line 234: The authors could consider adding an equation for the PWConv as well, next to Equation 2.*

Thank you again, we will implement these suggestions in our manuscript.

*Lines 253-255: Why did the authors choose to implement two kernels, and effectively employ a GNN with two types of nodes? Why not opt for a single node type (the stations), using river segment length and elevation difference as edge features between station nodes? A GRU could still be implemented, before, after, or during the message-passing iterations. I am wondering what motivated the authors to adopt this specific (and rather complex) model architecture? Was a simpler setup unable to achieve the desired results, or was there a conscious design decision behind this choice? I suggest elaborating on this part of the model architecture.*

The recurrent graph neural network is based on previous work from our lab, a model called DISTANA (Karlbauer, Matthias, et al. "A distributed neural network architecture for robust non-linear spatio-temporal prediction." arXiv preprint arXiv:1912.11141 (2019)). We will add a reference to this work in our manuscript. The idea is to have transition kernels whose main job is to aggregate and model the exchange of information between nodes, while the station kernels maintain information at the nodes themselves. In general, it is perfectly reasonable to try other kinds of graph neural networks here. We decided to use one that we already established before.

*Line 261: Initially, two types of kernels are introduced: station and segment kernels. However, line 261 refers to a "transition kernel". Based on Figure 2, I assume this is the same as the segment kernel. If so, I suggest consistently using the term "segment kernel" to avoid confusion. If my assumption is incorrect, then I suggest clarifying what is meant by "transition kernel".*

Thank you, this was relict from renaming the components during the writing process..

*Lines 261-262, "Each kernel first concatenates its static, dynamic, and lateral inputs. . . ": Based on Figure 2, it appears that the segment kernel only receives static inputs (adjacency matrix, elevation difference) and lateral inputs (from neighbouring stations). While these lateral inputs are dynamic in nature, the phrasing suggests a distinction between "dynamic" and "lateral" inputs. If "dynamic inputs" refer specifically to discharge and embedded runoff , then only the station kernels receive them. I suggest clarifying this distinction or rephrasing to avoid confusion.*

Here, we described the architecture in a more general way than we actually used it in our experiments. We will clarify this.

*Figure 2: I really like this figure, especially how the rainfall-runoff component is visualized and the connection between the two models is illustrated. However, I suggest adding more detail on the types and dimensions of the data streams, as these are not always clear from the text. For example, between the linear block and the StationGRU, you could add the label "Embedded runoff" and indicate its dimensions (e.g., Hx,y,c). I also recommend clarifying that the discharge data in the bottom left is observed discharge, while the discharge in the bottom right is simulated discharge. It would also be helpful to include dimensions for all input data, such as solar radiation, precipitation, altitude difference, river segment lengths, and the adjacency matrix.*

Thank you, this will certainly make the figure more comprehensible.

*Line 264, "Afterward, the tensor is split into dynamic and lateral outputs.": I suggest clarifying that the dynamic outputs refer to the estimated discharges, while the lateral outputs are embeddings used to update the SegmentGRU again, as indicated by the two outgoing arrows from the StationGRU in Figure 2.*

*Line 291: I suggest that the authors specify that they refer to a NVIDIA A100.*

*Lines 292-297: I suggest the authors provide further explanation regarding their choice of hyperparameters. For instance, were other optimizers considered besides Ranger? Did they explore the use of learning rate schedulers? What criteria were used to determine the sizes of the hidden states? Additionally, why was the SiLU activation function selected?*

*It is no problem if these decisions were made without any detailed exploration of alternatives, as the focus may have been on demonstrating the proof of concept rather than exhaustive model optimization. However, some more information about these choices (or the lack thereof) could help future researchers in determining what works well and what has not been tried yet (and thus where there*

*is potentially still room for improvement).*

Most of the hyperparamters were found experimentally. Indeed, we tried different optimizers, played around with learning rate scheduling, and tried different activation functions. Currently, we are still investigating the effect of different hidden sizes. In general, if two options produced similar results, we settled for the simpler one. Learning rate scheduling, e.g., did not improve the results and was therefore removed.

*Lines 326-327: I am unsure whether the inclusion of four different metrics adds value to the study, particularly if they are conceptually similar. For example, KGE consists of three components, one of which is PCC, and is essentially the same as NSE but without a certain bias. I suggest either replacing NSE and PCC with other, more distinct metrics or omitting them entirely.*

Thank you, we agree and will happily remove the NSE and PCC from the manuscript.

*4. Results*

*The results are described and illustrated in detail. However, aside from the four (relatively similar) metrics, the performance of DRRAiNN is not compared to that of EFAS in any other way. Additionally, the inferred contributing area per discharge station, as simulated by the DRRAiNN model, is compared to the catchment area delineated from the DEM. As the authors correctly point out in the introduction, this DEM-based delineation can be problematic, particularly due to its disregard for underground flows.*

*Nevertheless, a mismatch between the DEM-delineated catchment and the contributing areas from DRRAiNN is sometimes framed as an error by DRRAiNN (e.g., lines 420, 425). As the authors correctly note (lines 426-429), this mismatch does not necessarily indicate an error, but could reflect an unobserved subsurface flow path. I suggest presenting both the DEM-delineated catchment and the contributing area from DRRAiNN as feasible options, rather than assessing one against the other.*

We agree, this could be phrased better. When we say that the results are not perfect, we mean the fact that in some cases, a square is visible which corresponds to the receptive field of the CNN. We will describe this in more detail and consider both as reasonable options.

*If the authors wish to include a more objective assessment of the physical plausibility of the DRRAiNN model, they could train the model on EFAS-simulated (or any other PBM model) discharge instead of observed discharge. Since the contributing area is known a priori in this case (the DEM-delineated catchment), training DRRAiNN on EFAS would ideally result in a perfect alignment between*

*the contributing area according to DRRAiNN and the DEM-derived catchment. While such experiments may be outside the scope of the current manuscript, they could be explored in a follow-up study.*

*Furthermore, I had hoped to see in this chapter some analysis of the internal system states, especially given their limited number (either 4 or 8) and they're location in GRUs or LSTMs. It's conceivable that certain hidden states respond to specific inputs, which could make them at least partially physically interpretable. While I don't expect this to be included in the current study, it's a valuable direction the authors could consider for future research.*

Both of these are very interesting endeavors that we considered but didn't find the time yet to pursue. We will make sure to mention them in the future work.

*Lines 369-371: "We likely chose ... considered stations." I suggest leaving this line out, as in my opinion it is redundant.*

Agreed.

*Figure 5: Nice figure! I suggest including the unit ($m^3/s$) in the MAE plot. Also, if I understand correctly, the DRRAiNN performance is based on the three best-performing seeds. I would suggest explicitly mentioning this in the caption.*

*The large difference in performance across different seeds is interesting. With better learning rate scheduling, this issue could potentially be mitigated or at least reduced. Bentivoglio et al. (2023, see https://doi.org/10.5194/hess-27-4227-2023) employed a curriculum learning strategy, similar to the TBPTT used here, and found that gradually lowering the learning rate every few epochs improved performance.*

Thank you, we tried learning rate scheduling in the past but didn't not find a significant difference in performance (or spread thereof). However, currently we running some promising experiments that do not seem to suffer from this. We might be able to put these new results into an updated version of our manuscript.

*Line 406: In lines 442-444, the authors (correctly) list some factors that could contribute to the difficulty in predicting streamflow at certain stations beside unobserved subsurface flows, but these are not listed here. I suggest listing these also (or only) here.*

Referee comment RC1 suggested that an analysis of these factors would be interesting, with which we agree. As this is out of the scope of this study, though, we will most likely keep them in the future work section and remove them from the results section.

*Figure 6: I find the information in this figure very interesting, but the figure*

*itself could be improved. While this is subjective, the figure might look more visually appealing if the performance at each station were represented as markers rather than by a line connecting the performance per station. The range of DRRAiNN performance across different seeds could be indicated with whiskers. Additionally, the mean discharge per station could be plotted as a dashed line behind the performance markers for better visibility of the markers.*

Yes, in retrospect it makes no sense to connect the stations. We will update the figure accordingly.

*Figure 7: This is a very interesting figure and, in my opinion, showcases one of the most important contributions of this study. However, I believe the figure could be improved in several ways. First, the individual plots are quite small and difficult to read; I suggest making them larger. The DEM-based catchment area is barely visible. Additionally, displaying the same legend in all four plots is in my opinion unnecessary; I recommend having a single legend, placed outside the plots.*

*The station name is already in the title of each plot, so repeating it inside the plot is redundant and makes the plot harder to read. Instead, I suggest indicating the station of interest with a different marker.*

*Furthermore, though this is subjective, I believe the plots would benefit from being zoomed in on the station and its catchment area. There is no need to display the entire river network, as the goal of this figure is comparing the DEM-derived catchment and the attribution area from the DRRAiNN model. Therefore, it would be sufficient to zoom in per station and show the station, the DEM-derived catchment area, the attribution area, and possibly the river network within the catchment.*

*Finally, as an ambitious suggestion (and possibly challenging to implement in 2D), it would be fascinating to see the attribution overlaid with the DEM itself. This would provide insight into how the area identified by DRRAiNN as significant aligns with the DEM.*

Again we agree, the plots can be improved substantially. Thank you for the concrete feedback, we will surely make use of it.

*5. Discussion*

*I recommend expanding the discussion to address whether and how future work could overcome the limitation of estimating discharge only at observation stations. More importantly, I suggest to discuss whether the model could even become fully independent from discharge observations, which would expand its applicability to ungauged basins. These are two important drawbacks of the current model, but neither are addressed explicitly.*

We discuss a little bit of spatial generalization, but agree that this could be extended. In some preliminary experiments, we simply treat individual stations from the training data set as "virtual" stations. This means that these stations are part of the graph, but the model never receives discharge as input or target at these stations. It seems that the network can predict discharge at these locations to some extent already. Another option would be to remove the graph neural network completely from the model. However, some form of target data will always be required. These must not necessarily come from gauging stations, though, as there are efforts to estimate discharge from satellite observations. We will make sure to discuss these aspects in our future work section.

*Furthermore, I suggest reordering the discussion somewhat as to group related paragraphs together. E.g. lines 485 – 489 discuss including more input data, and line 494-499 discuss including more output data.*

*Additionally, the overall discussion is quite long, and I suggest trying to condense some paragraphs. Line 447-453 could maybe be moved to the chapter 4.*

*Lines 462-470: In my opinion, this should be part of section 3.2.*

These suggestions are in line with Referee Comment RC1. In a future version of the manuscript, we will try to condense the discussion, especially the future work.

*Line 489: A warm-up period of 10 days is quite an achievement! Most models, especially large and complex PBMs, often require warm-up periods of a year or longer. A model with such a short warm-up period is impressive, and I believe the authors should highlight this accomplishment more explicitly.*

Thank you! We were not aware of that and are happy to highlight this in our discussion.

*Lines 491-492: Assuming the authors are referring to remotely sensed soil moisture, this would be limited to the moisture in the upper soil layers, not the deeper, saturated layers. Therefore, I believe the initial 10-day warm-up period is still necessary, even with soil moisture as an input, since the model will also need time to adjust the deeper soil moisture. However, including compressed meteorological data might proof effective.*

Again a valid point that we were not aware of. We might remove this point from the discussion in an effort to condense.

*6. Conclusion*

*The conclusion is well written and concise.*

Thank you!

*Appendix A*

*This is overall a very nice addition to the study, and similarly well written.*

*Figure A1: To be consistent with Figure 5 and Figure 6, I suggest to plot the data from EFAS again in orange.*

*Figure A2: This is very interesting, and I believe plotting the entire region is now warranted, as the attributed cells are located far outside the DEM-derived catchment. However, I still suggest increasing the figure size, using a single legend for all four plots placed outside the plots, and differentiating the stations with distinct markers rather than displaying their names.*

*Lines 543-544: If I understand correctly, the entire ConvNeXt block has been removed from the architecture. I suggest explicitly stating this for clarity.*

*Figure A4 and Figure A7: same comment as for Figure A1.*

*Figure A5 and Figure A8: same comment as for Figure A2.*

*In the ablation studies described in A2 and A3, significant portions of the model are removed, which likely results in fewer model parameters and faster runtimes. I suggest discussing the model size and computational efficiency in this context as well.*

We will certainly improve the quality of our figures, including the ones in the appendix. You are correct, the model has fewer parameters and different runtimes. We will look those up and report them in our revision.

*Technical corrections*

*Line 290: Missing "table".*

*Figure 4: Missing "discharge" after "... and highest (d)".*

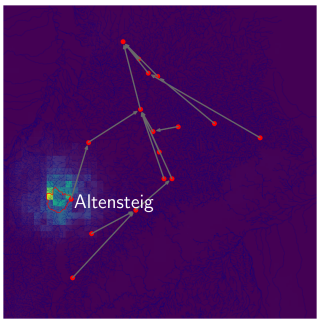
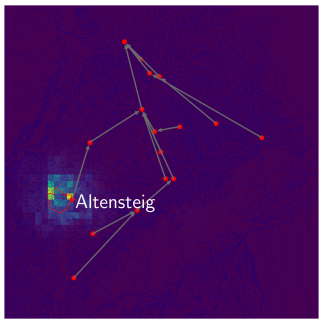
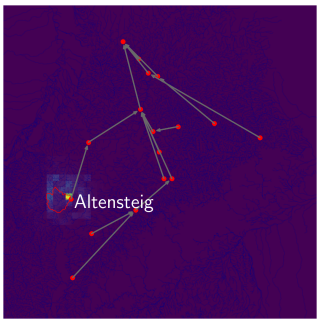
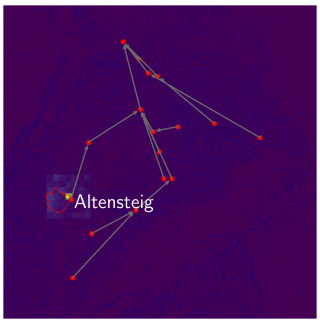
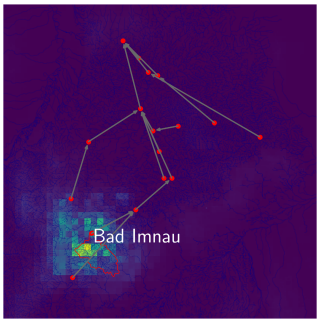
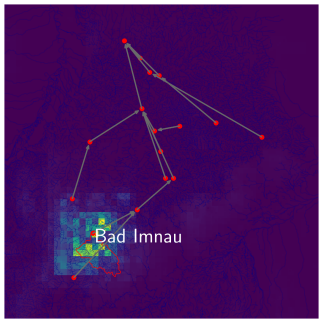
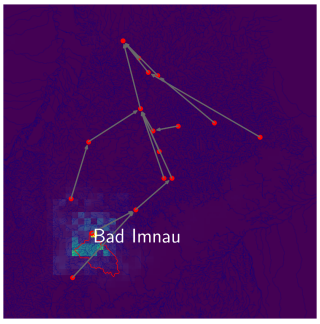
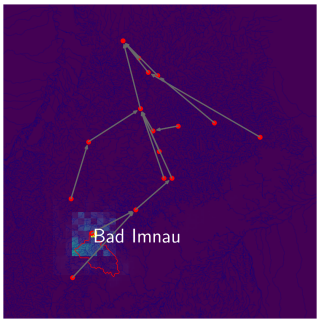
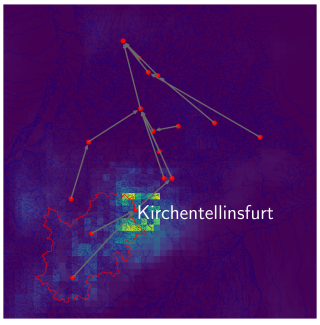
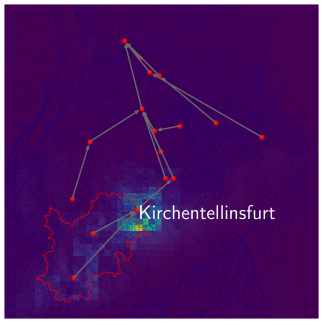
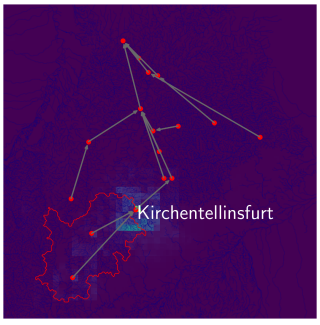
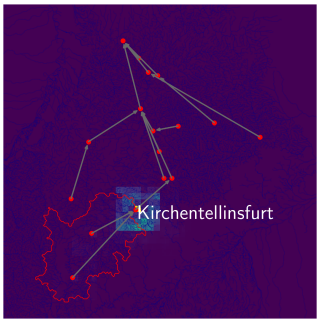
Thank you!

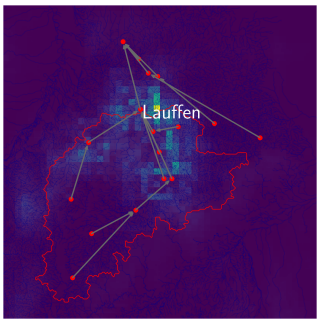
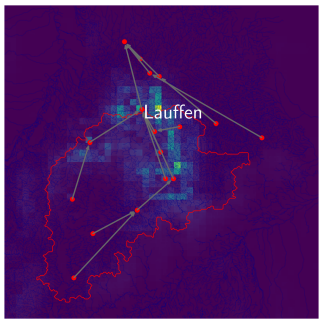
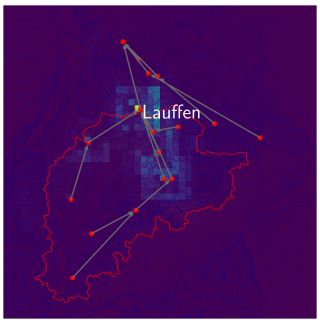
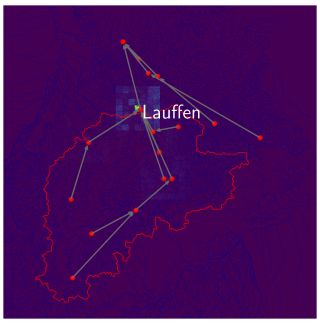
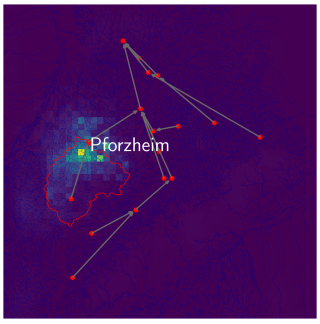
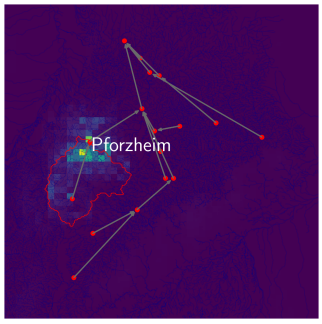
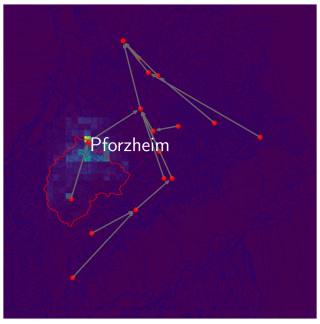
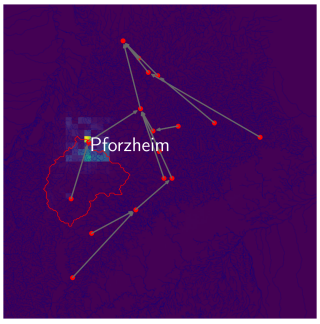
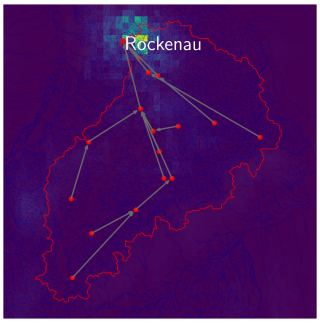
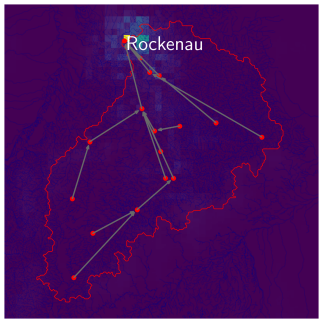
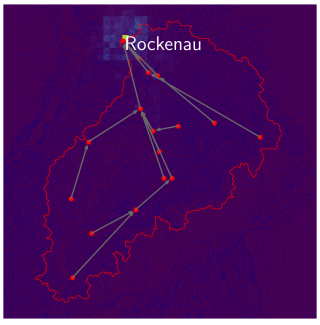
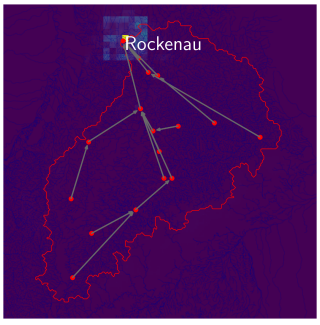
The following figures show attribution maps for different stations depending on how far we look into the past.

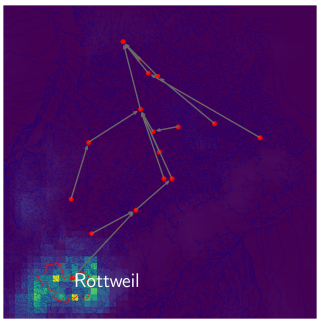
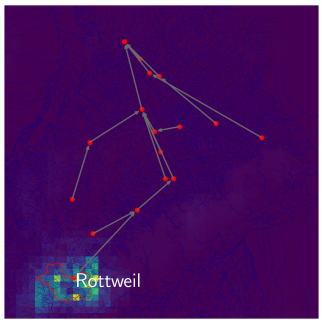
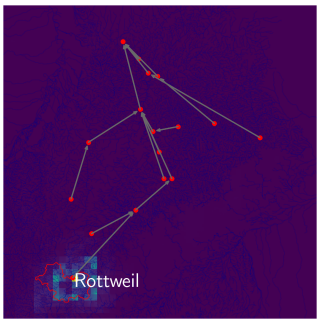
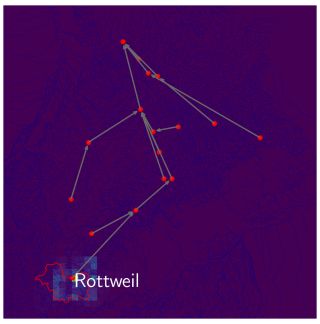
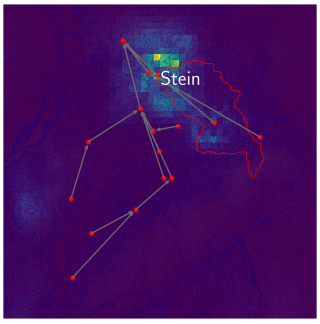
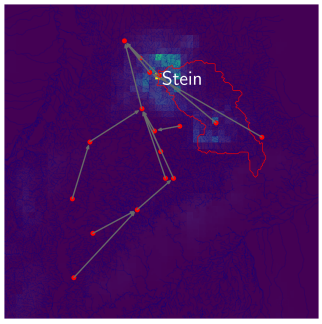
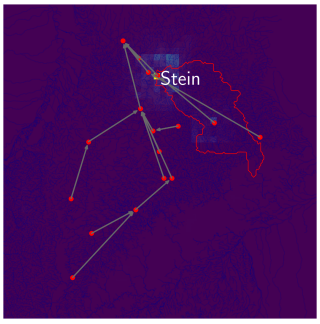
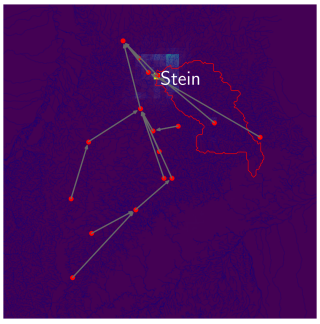
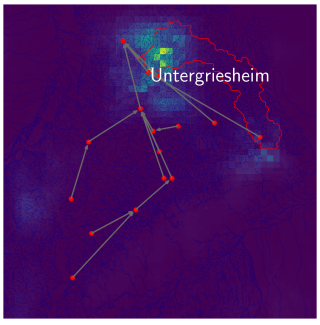
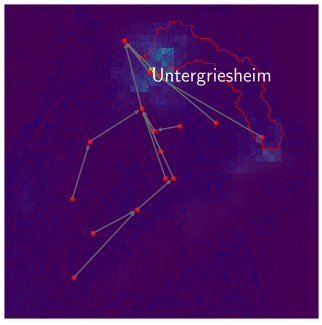
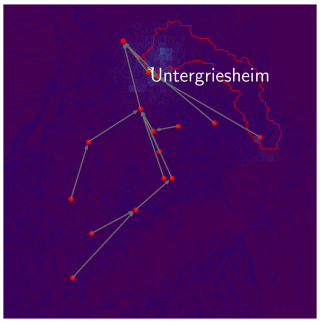
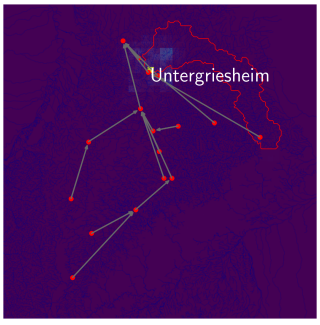
They were created as follows: For each station, we feed all 20-day sequences from our validation dataset through the model. Afterwards, we compute the gradient of the discharge output in the last time step with respect to the precipitation input in each sequence, grid cell, and time step, and multiply it with the precipitation itself. We take the mean over the validation dataset and split the time dimension into 4 subsequences: days 0-4, 5-9, 10-14, and 19-20. For each subsequence, we again take the mean of the attributions. The results are depicted below with the most left figure depicting the attributions from the initial 5 days and the right most figure showing the attributions from the most recent 5 days.

The further we look into the past (i.e., the further we look to the left), the larger is the area from wich precipitation contributed to the discharge prediction. We take this as evidence that our model actually propagates quantities over space and time and physically reasonable manner, at least to a certain extent.

It should be noted that, again, these results are not 100% consistent. We only show the results from the best seed here. Overall, three of our five seeds show produce reasonable attribution maps.


Altensteig

Altensteig

Altensteig

Altensteig

Bad Imnau

Bad Imnau

Bad Imnau

Bad Imnau

Kirchentellinsfurt

Kirchentellinsfurt

Kirchentellinsfurt

Kirchentellinsfurt


Lauffen

Lauffen

Lauffen

Lauffen

Pforzheim

Pforzheim

Pforzheim

Pforzheim

Rockenau

Rockenau

Rockenau

Rockenau


Rottweil

Rottweil

Rottweil

Rottweil

Stein

Stein

Stein

Stein

Untergriesheim

Untergriesheim

Untergriesheim

Untergriesheim