The revised manuscript is an improvement over the original. Some of the language used in asserting likelihood has been improved, the scope is more appropriate (e.g. delineation vs localization), and interesting new analyses have been performed. While details were added to the introduction, more are required as the background (especially on model interpretability) is still insufficient. Overall, the science is interesting but persistent imprecise and non-specific language remains as a substantial barrier.

The new analyses performed (varying the catchment location within the input domain) provides very exciting results. While many studies exist that build a machine learning model for streamflow prediction, I think that this sensitivity analysis makes the study stand out. Figure 7 shows very promising results and strongly supports the authors' case. I was very happy to see this. I think that this result is interesting, and it is worth the effort to improve the quality of the rest of the paper to continue towards publication. I think that this result will generate interest in the community.

Unfortunately, the writing is frequently imprecise and unclear. There were many instances where language surrounding precision was improved from the initial submission to this submission; however, many instances remain. This is a problem because vague phrases can be interpreted in multiple different ways, some of which are incorrect. These instances act as a barrier to readers' understanding. This is a substantial problem, but since the science seems to be sound and interesting, it is a problem that can (and should) be overcome. In this review I note many examples where improvements can be made, and why. It is my hope that the authors take the principles behind these comments and apply them throughout the text.

Comments:

Line 2: An ANN does not necessarily offer a "convenient solution" – perhaps "potential solution"?

Line 2: Input-output relationships are typically not "simple"

Line 4: "…few climate stations within a karst spring catchment are available" – it is not clear as to what an 'available climate station' means. Are there stations within these catchments but they are 'unavailable' for use? Or are they not located within these catchments? An alternative: e.g. "…few climate stations are located within or near karst spring catchments".

Line 4: "Hence spatial coverage… severe uncertainties." Introduce uncertainty into what?

Line 9: "based on" is not precise – the authors mean "use climate station data as input".

Line 10: The authors qualify their results as being "excellently suited to model karst spring discharge". This should be quantified with specific results (e.g. NSE or another quantitative

metric); otherwise, this abstract is not actually giving the reader the specific results of the study.

Line 11: This sentence is a bit hard to follow and can be made more clear if split in two, e.g. "The 2D-models show a better fit than the 1D-models in two of three cases and automatically learn to focus on the relevant areas of the input domain. By performing a spatial input sensitivity analysis…".

Line 17: "High heterogeneity" of what?

Line 19: Terms like "of them" are vague. What specifically does "them" refer to?

Line 19: "Certain level of background knowledge about the system" – All models, physically based or empirically based, require a 'certain level' of background knowledge (using an ANN still requires selecting input data/domain/etc). Do the authors mean that physically based models require more detailed system knowledge?

Line 20: "contrary" → should be "contrast"

Line 22: The sentence beginning with "Even though" seems to have two contradictory statements: first, that ANNs are not standard practice, and second, that they have been used for quite a long time. So what is it that the authors are trying to say here?

Line 23: The sentence beginning with "In fact" (and throughout the text) should be improved by using an active voice. E.g. "Johannet et al. (1994) showed that modelling water infiltration in karstic aquifers is possible with ANNs and was one of the first applications of ANNs in water related research."

Line 25: Sentence starting with "The number of applications" is not very clear (what does "even more accelerating" mean?). An alternative: "Applications of ANNs in hydrology have received a substantial amount of research attention, for example for rainfall-runoff modeling (e.g. Maier and Dandy (2000), Maier et al. (2010)), for groundwater applications (e.g. Rajaee et al. (2019)) and for hydrology and water resources in general (e.g. Sit et al. (2020))."

Line 28: Here, a transition from ANNs to CNNs should be included. What are the limitations of the fully-connected ANNs described in these reviews (e.g. MLP models)? Why not use them here? What are the characteristics of CNNs? Why are they suitable for application here? Then state that you use CNNs which wave been successfully applied etc etc. This will help improve the logical flow of the introduction: Karst systems are complicated; fully-connected ANNs are good for predictive modelling in hydrology; ANNs are limited by x y z; CNNs are beneficial because of a b c; so we use CNNs here to address the challenges in karst modelling.

Line 34: "Rudimentary" is a negative descriptor. Best to remove.

Line 35: “Especially… results,” can be removed, unless the authors wish to get into conceptual models in more detail.

Line 39: “For this… catchment either.” These two sentences have several examples of what I mean when I mention imprecision limiting reader understanding. “For this” – for what? “Reasonable time periods” – what is reasonable? “Various products” – like what, specifically? “Extract corresponding time series” – time series of what? “Knowledge of the spring catchment” – what kind of knowledge? “The available grid cells do not exactly match the catchment either” – match in what way? What does “exactly match” mean? The total lack of specificity greatly limits the informative content of these sentences.

Lines 42 – 49: These sentences are about catchment delineation. I think this could be a new paragraph since it does not really connect with the ideas of spatio-temporal completeness of the prior sentences. Since the goal of the paper has moved from delineation to localization, this should be reflected here as well; e.g. the ideas can flow from why delineation is a goal, how it is done, to why localization is a goal, to how it is done, to then how the authors propose to do it.

Line 54: “by themselves” – I think this should not be used as a descriptor for what the models are doing. Perhaps “automatically learn” is a more accurate way to say this.

Line 55: “… 2D-CNN processes … time series.” This phrase is an example of when unclear language can imply an incorrect understanding. A reader could interpret this as meaning that the 2DCNN and 1DCNN are entirely independent of one another, when really they are not (e.g. both components are used for forecasting of the spring discharge time series). The 2D CNN learns spatial features and the 1D CNN learns the temporal relationship between those spatial features – perhaps that’s what the authors mean here.

Line 64: I still have trouble with this phrasing. I think it is presumptuous to expect ANNs to be “superior”. I think it is better to phrase as a potential advantage, e.g. "These requirements hold for our proposed methodology as well, but a potential advantage of ANNs is their nonlinearity which may better capture the nonlinear relationships between rainfall and discharge."

Introduction general comment: There should be information about deep learning model interpretability. It currently reads as though the authors found an interesting study and modified their method, without consideration for any other methods. There should be more information given (e.g. see studies like McGovern et al (2019), Fleming et al (2021a), and Fleming et al (2021b) and the references cited therein).

Line 79: What does "and except Unica mean"? Data was not easily accessible? Or that previous modelling approaches are not available for comparison?

Line 93: What does “to smaller parts” mean?

Line 124: What is meant here again by “parts”? ‘from A and C (Fig. 1b)’

Line 124: “Clearly of the channel flow type” – Remove the word “clearly”.

Line 164: “Including pumping data…” – Better to use active voice: “We do not use pumping data as input in this study because…”

Line 169: “They use a wide variation…”: This is vague and doesn't offer any specific information. What are the input data used and what are the study goals?

Line 172: Why are these studies best compared to the authors’ work?

Line 189: “Smallest” and “largest” are modifying "the relation of grid resolution to catchment size"; what relation? The ratio? Do the authors mean to say that Aubach is the smallest catchment in terms of area, or that it has the smallest ratio of grid cell area to catchment area?

Line 195: I don’t understand what the authors are saying with “where the RADOLAN grid cell center lies in”. What does that mean?

Line 206: This claim can be made stronger by citing a study which quantifies the performance of at least the most important metric (precipitation). There are many reanalysis products available. Why did the authors choose the ones they chose?

Line 213: Provide references for CNNs applied for object recognition, image classification, and signal processing.

Line 218: "Density of information" isn't a phrase that makes sense to me.

Line 222: As a reader, I am not going to trust statements like “from our experience in preliminary work”. Either provide results or do not include statements like this.

Line 224: "We have shown they are faster" -- seems to be in contrast to the statement in the response to referees, where the authors claimed that "there is no large difference in training time". When the authors say they “have shown” – where? Where is it shown that CNNs are faster than LSTMs? The authors have made this claim but do not provide evidence.

Figure 2: Since the authors describe the input data difference on line 229, this Figure can be improved by making this distinction clear visually as well (e.g. label with weather station vs gridded reanalysis data input).

Line 231: See earlier comment about possible misinterpretation of this phrasing.

Line 232: I don't I agree with the statement that a 1D CNN instead of an LSTM improves the comparability between the two modelling approaches (2D-1D CNN vs 1D CNN). In my opinion

those are just as comparable as a 2DCNN-LSTM vs LSTM. I think this sentence and argument can be removed.

Line 261: “Sinus” – the authors mean “sine” or “sinusoid”

Line 272: “… physical meaning… others.” This phrase is confusing to me -- alternatively can say "CNN-LSTM models can learn to focus on specific areas of the spatially distributed input data".

Line 272: The sentence beginning with “We modify…” doesn't make sense ("... we demonstrate that this approach to qualitatively localize..."?) I think some text was inadvertently deleted here.

Line 278: "Absolute parameter value of each pixel" --> "Absolute value of each variable at each pixel"

Line 282: Here, and throughout the remaining text, the authors refer to error but it is not clear what they mean. How is error defined? Difference between perturbed prediction and observation or between perturbed prediction and unperturbed prediction?

Line 282: “Repetitions” – I believe the authors mean “iterations”.

Line 286: The sentence beginning with “Also these” is not easy to follow. It should be rephrased to be more clear (e.g. “Temperature is spatially autocorrelated, meaning that temperature information from outside the catchment area may be used to infer temperature within the catchment area. In contrast, precipitation is less spatially autocorrelated, meaning that precipitation information from outside the catchment area is less related to precipitation from inside the catchment area. Therefore, we hypothesize that the within-catchment precipitation fields will be most important for the model’s prediction, and we will test this hypothesis by visually inspecting the sensitivity maps produced by the modified approach of Anderson and Radic (2021).”).

Line 301: “Daily” – I think the authors mean “diurnal”.

Line 307: The authors claim that diurnal oscillations “no longer appear” in summer and autumn. They are still visible in Fig 3b. They are diminished, but still appear.

Line 312: Here, and throughout the text, the authors refer to uncertainty but it is not clear what they mean. Do they mean error between prediction and observation? This should be clarified throughout the text.

Line 322: What range? Instead of “same range” I think it is better to say “Our model has a similar, albeit slightly lower, NSE value compared to these three models. One reason for this discrepancy could be that none of the previous studies…”.

Line 338: The sentence starting with “Additionally” is unnecessarily complicated. It can be clarified by stating “The optimized model uses P, T, …”.

Line 347: Sentence starting with “In total” – where do “we see” this? Where is it shown? And again, what is meant by “uncertainties in terms of input data”?

Line 368: It is more accurate to say “good” rather than “solid” (this occurs other times as well).

Line 368: What do the authors mean by “reaction”? Are they referring to the streamflow response (e.g. where observations increase strongly)?

Line 369: What is meant by “conservative”? An underestimation?

Line 374: What is meant by “even more true”? How can something be more true than observations?

Line 376 – 382: Could this conjecture not be easily verified by looking at the training predictions and seeing if these same errors persist? (E.g. underestimating spring peaks and overestimating low flows even before 2014)

Line 386: “by many orders of magnitude” – It looks like flow in Fig 2 varies by 2 orders of magnitude, not ‘many’. Be specific.

Line 389: That low flows are no longer overestimated by very much seems to imply that the authors' previous claims about the impact of land use change may be incorrect

Line 392: What “same conceptual understanding” are the authors referring to?

Line 395: If there are some precipitation events in the gridded data that are not in the station data, wouldn’t there be additional modelled discharge peaks in Fig 4b as compared to Fig 4a? The authors should be clearer in pointing to evidence to support their claim here.

Line 406 – 409: These studies should be moved to the introduction. Why are they in the discussion if they are not discussed? What did they do? What were the goals?

Line 419: Remove “easily” and provide a reference linking RH and radiation to evaporation.

Line 419: Remove “basically”.

Line 425: Remove “definitely”.

Line 426: Sentence beginning in “Though”– what does it mean? It is not clear.

Line 439: What does “knowledge extraction” mean? Be more specific.

Line 441: What does a “larger database” refer to? More observations for training? More input variables? Be specific.

Line 454: “hardly the size of one grid cell” – just say “smaller than one grid cell”.

Line 456: The phrase “main direction of the weather area” is not clear to me. Upwind? Justify why it is the “main direction”. This phrase is used multiple times.

Line 458: “… this effect should be related to the size of the filter” – can the authors explain why and/or provide a reference for this?

Line 475: Remove “real world”

Line 476: Remove “even though… discharge signal.” It doesn’t add to the discussion.

Line 491: Support point (i) with a reference and explain/clarify what is meant by “lower dampening” in point (ii).

Line 498: Given *what specifically* about the spatial resolution, heatmaps, and simulation results makes Unica springs the best example to investigate? Make your thought process *explicit*, otherwise it sounds like the authors tried all three basins and are cherry-picking the results.

Line 501: “data frame” (and throughout the remaining text) – This term (and ‘dataframe’) is well used to refer to a type of data structure. I believe the authors mean ‘domain’ or a similar term.

Line 504: What is meant by “extract the relevant input data” and where is this shown? From the heat maps, or from the predicted discharge?

Line 537: It is not shown or explained how/why the 2D approach reduces the amount of work.

Line 543: What inaccuracies? Be explicit and specific.

Line 545: “…we assume it can be used to delineate catchments quite accurately” – this is not a conclusion and is not really supported by the study in its current form.

Line 548: Hard to conclude that 2D is overall superior due to the performance metrics -- maybe could state something like: "A key benefit of the 2D approach, which uses spatially discretized input data from climate reanalysis, is the spatially and temporally complete nature of the data and the number of variables available for study" or something to that effect.

Line 549: Sentence starting with “though”: Increased effort for what as compared to what?

Table D1: Could the input sequence length be related to features observed in the predicted streamflow? E.g. diurnal oscillations are modelled in August seems it could be due to the model mapping temperature to flow, but due to the input time series length the model may not necessarily knowing if there is a snowpack available for melt (since this accumulated on longer timescales than are provided as input)? This point is worth considering and potentially adding to the discussion.

References

Fleming, S. W., Vesselinov, V. v, and Goodbody, A. G. (2021). Augmenting geophysical interpretation of data-driven operational water supply forecast modeling for a western US river using a hybrid machine learning approach. Journal of Hydrology 597, 126327. doi:https://doi.org/10.1016/j.jhydrol.2021.126327.

Fleming, S. W., Garen, D. C., Goodbody, A. G., McCarthy, C. S., and Landers, L. C. (2021). Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence, 602, 126782, https://doi.org/https://doi.org/10.1016/j.jhydrol.2021.126782.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., et al. (2019). Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. Bulletin of the American Meteorological Society 100, 2175–2199. doi:10.1175/BAMS-D-18-0195.1.