Response Letter

We thank the referee for this extremely detailed textual analysis and the positive judgement of the manuscript in general. The propositions made by the referee considerably helped to improve the quality of the manuscript. We hence were able to improve the language in general and hope, that we are more precise throughout the text. Please find our answers in the following in red, while the review comments are reproduced verbatim in black. Line numbers in red, refer to the revised version of the manuscript. Some paragraphs were substantially rewritten, we therefore do not provide a Line number for every answer. The tracked-changes document hopefully helps out in these cases.

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"?

We agree. We rephrased the respective part. Lines 1-3

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

We rephrased this part, please see the answer above. Lines 1-3

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".

We have adopted the suggestion. Thank you. Line 4

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

Sentence revised to: "Hence, spatial coverage is often not satisfactory and can result in substantial uncertainties about the true conditions in the catchment, leading to lower model performance." Lines 5f.

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

Correct, thank you. Sentence adapted to: “We compare the proposed approach both to existing modeling studies in these regions and to own 1D-CNN models that are conventionally trained with climate station input data.” Lines 9-10

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.

Thank you for pointing out. We added NSE and KGE ranges to the Abstract. Line 11

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…”.

Thank you very much. We have adopted the proposed modification as it stands. Lines 12-14

Line 17: “High heterogeneity” of what?

Karst systems in general are characterized by high structural heterogeneity due to the at least in large parts unknown conduit network, which controls the highly variable groundwater flow. We modified the respective sentence to clarify. Lines 17f.

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

We revised this section considerably. Please see Lines 18ff.

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?

We revised this section considerably. See especially Line 22

Line 20: “contrary” à should be “contrast”

Thank you for pointing out. We revised our formulation, which is also now embedded in a considerably reworked paragraph. Line 21

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?

In our opinion this is indeed contradictory. Being a useful tool (stated before) and at the same time being used for a long time already (stated here), implies that ANNs may be an established approach. This is not the case. We therefore stick to this specific formulation.

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."

Done. Thank you for pointing out. Line 25f and throughout the manuscript.

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)).”

Thanks for pointing out. We restructured these sentences and added more background information as requested below. Lines 26ff.

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.

Done. Thank you for pointing out. See roughly Lines 25 to 44

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

Done.

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

Done.

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.

Thank you for pointing out this specific example. It certainly helped to get your point. We revised this sentence and others throughout the text. e.g. Lines 65-68

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.

Thanks for pointing out. We have revised this section. Lines 74-84

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.

Changed throughout the manuscript, thank you.

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.

Again, thanks for pointing out. We rephrased these (and other related) sentences to become clearer. See e.g. lines 85ff or section 3.2

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."

Thank you again for your effort to propose adequate changes. We adapt this one to our text verbatim. Lines 99-100

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).

Thank you for improving also our storyline. We added a paragraph on explainable AI and model interpretability to the introduction. Lines 45ff.

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

We made the reference clearer and revised the respective sentences. Esp. L 121f. but generally lines 101 ff.

Line 93: What does "to smaller parts" mean?

We change the wording ("proportions") and hope it gets clearer now. L150

Line 124: What is meant here again by "parts"? 'from A and C (Fig. 1b)'

Parts as indicated on the map. We now write "regions". L182

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

Done. L182

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

Done. L222

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?

We revised the literature discussion throughout the manuscript. Thus, this paragraph was completely rephrased and moved to the Introduction section Lines 101ff

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

We revised the literature discussion throughout the manuscript, and now better explain these aspects. The paragraph was completely rephrased and moved to the Introduction section. Lines 101ff

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?

We revised the wording in the manuscript and now speak of “ratio”, however this specific sentence was removed during rewriting the corresponding section.

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

We rephrased it to: “The higher resolved precipitation data from RADOLAN is thus augmented with climate parameter values from ERA5-Land, which were downscaled and re-gridded to match the 1x1 km RADOLAN grid.”

We hope this answers the question. L245f.

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?

We added a study that evaluates precipitation for both data sets and we better explain why we chose them. L259

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

Done. L266f.

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

We have revised and augmented this passage: “The latter performs down-sampling of the produced feature maps and summarizes the features detected in the input. This decreases the total number of parameters of the model and makes it approximately invariant to small translations of the input (Goodfellow et al. 20116)” Lines 271f

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.

We removed this part of the respective statement and now only refer to the cited studies.

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.

After stating that CNNs are faster we now cite Wunsch et al. (2021), where we investigated this aspect in the context of groundwater level forecasting and showed that CNNs are systematically faster than LSTMs. In our last response to referees, we referred to the difference between 2DCNN-1DCNN and 2DCNN-LSTM, whereas we here speak of 1D-CNNs. We have additionally reworked the whole paragraph (including additional references), to point out that the speed is only one aspect in our choice for CNN instead of LSTM models. Lines 277ff.

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).

Thank you for this proposition. We modified the Figure accordingly.

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

We similarly rewrote this whole paragraph. See also our answer to Line 224

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.

We think that using 2DCNN-1DCNN instead of 2DCNN-LSTM allow better conclusions on the importance of the input data and its influence on the modeling result. Using a LSTM would mix the data influence with the question whether the model performance changed due to the LSTM. We elaborate this aspect now in the manuscript. Thanks for pointing out. Lines 287-289

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

Yes, thank you. A translation error from our side.

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".

We adapted the proposed change. Thank you. Lines 325f.

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.

You are right, thanks for noticing. We now go with: “We modify this approach and transfer it to karst spring modeling, where we demonstrate that this approach is suited to qualitatively localize karst catchment locations. “Lines 326f.

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

We adapted the proposed change. Thank you. L330

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?

It’s the difference between the perturbed prediction and the unperturbed prediction. Thank you for pointing out this missing detail, which we now mention in the text. Lines 334-335

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

Corrected, thank you.

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).”).

Thank you again for your excellent propositions. We rephrased these sentences accordingly. Lines 336ff.

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

Yes, thank you. We changed several occurrences throughout the text.

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.

You are correct. We changed it to: “now are diminished” L. 358

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.

We changed these occurrences throughout the manuscript. E.g. here we now use “error source”. L361f. Sie also our answer to “Line 347”

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…”.

Thank you for your excellent proposition. This sentence is, however, removed due to restructuring and rewriting the section. Lines 363-369

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

Thank you, we simplified the sentence to: “The optimized model uses P, T, Tsin, SMLT, SF, SWVL1/2/4 as inputs, thus omits E and SWVL3.” L.371

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”?

We rephrased this to: “In total, we think that both the 1D and the 2D-approach for this catchment bear substantial shortcomings in terms of how well the input data represents the true conditions in the catchment, even though the simulation results are generally very accurate” Lines: 392-394

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

Thank you. We changed several occurrences in the text. L404

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

Yes, thank you for pointing out. “Response” is the accurate wording. L405

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

Probably another mistake made inadvertently by transposing an additional meaning of a word into another language. Sorry for that. What we meant is that the slope is not as steep as for the observed recession. We rephrased the two occurrences in the text to be more precise. L 405 ff.

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

As we explain in the two sentences before, under certain conditions it is “impossible to accurately monitor the inflow conditions” This means at times of the plateau like peaks, we do not know the true conditions. Therefore, the simulated conditions might be conceptually true, thus representing a flow

variation we cannot observe. Nevertheless, we rephrased these sentences, to improve clarity. Lines 408ff

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)

Good point, thanks for pointing out. We think it is not that easy. As you can see in Table 1, we split the time series into four parts: 1981-2012 for Training, 2013+2014 for Validation (Early Stopping), 2015+2016 for HP Optimization and finally 2017+2018 for testing. The considered period of the environmental changes 2014-2018 is not part of the training data, but nevertheless has influence on the modeling process, since it is covered by the period used for HP optimization and early stopping. So even by looking at the different parts of the data it should be hard to disentangle these effects. Anyhow, we found it worth to notice. We rephrased this part: Lines 414ff.

Line 386: "by many orders of magnitude" – It looks like flow in Fig 2 varies by 2 orders of magnitude, not 'many'. Be specific.

Corrected. Thank you. L427

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?

Thank you for pointing out. We admit that we missed this flaw in our argumentation. We thus rephrased both parts, still mention the environmental change but do not hold it accountable for the 1D model performance during low flow. We also explain what we mean with same conceptual understanding. Lines 413ff. and Lines 433ff.

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.

You are right, we rewrote this sentence. L 436f.

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?

We moved these studies to the introduction. Thank you for pointing out. L. 106-107

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

Done. L45ff.

Line 419: Remove "basically".

Done.

Line 425: Remove "definitely".

Done.

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

We rephrased this section. L462ff.

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

We removed this sentence due to its missing relevance.

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

We added some details and moved this part to the introduction. L109ff.

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

Done. Thank you. L479

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.

Sorry for this imprecise description, presumably resulting from a translation problem. In German this literally describes the direction at a certain location from which weather phenomena usually originate most of the time. For example, when precipitation events usually come from the West, then this is what we mean. We acknowledge that this obviously is confusing. We decided to remove these statements from the text.

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?

We rephrased this statement. Lines 482-486

Line 475: Remove “real world”

Done, thank you.

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

Done.

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

We have removed this statement from our text.

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.

Done. L520ff

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.

Dataframe is also well used in GIS context to the frame on a map with two-dimensional content displayed. Nevertheless, to avoid misunderstandings, we have changed the wording throughout the text to “considered area of the input data”

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?

By “relevant input data” we mean “relevant grid cells within the considered input area”. We changed the wording accordingly. L527ff.

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

We do explain it now. Thanks for pointing out. L562ff.

Line 543: What inaccuracies? Be explicit and specific.

We rephrased the whole section. L566ff.

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.

We rephrased the whole section. L 572ff.

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.

We rephrased the respective sentences to make our reasoning clearer and also added the proposed sentence. Thank you. L574ff.

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

We rephrased this sentence. L578f.

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 be 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.

Thank you for this interesting remark. We provided Tsin as input variable to cope with this aspect. Using this sine signal input, the model (presumably) learns the season and the current position in the annual cycle. It may theoretically therefore be aware of a potential snow pack. You state that it might be due to the model mapping temperature to flow – we think that of course T is (somehow) mapped to flow, otherwise the model would not use it as an input variable. The question is rather if it was done in the correct way, such as deriving maybe seasonality from it, reducing flow in periods of high evapotranspiration and so on. To explore this aspect is an interesting idea, using explainable AI methods (e.g. SHAP values), to explore the influence of each input on the model output. This is however, beyond the scope of our study.

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/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: https://doi.org/10.1175/BAMS-D-18-0195.1.