We thank all Referees for reviewing our manuscript. We appreciate the constructive comments and will provide a point by point answer in the following. Please find the original comments in black and our answers in red. Line Numbers in our answers refer to the revised version of the manuscript.

Comments by the Editor:

Dear authors,

The two reviewers have completed the reviewing and pointed out several key limitations of this study. After carefully reading the manuscript, unfortunately, I have the similar opinions. Therefore, I expect that the authors take these comments seriously when they revise the manuscript, especially the follow comments:

1) A substantial revision of the introduction is needed by expanding to consider the further history of ANNs in water related research.
2) The choice of ANN model structure needs to be further justified.
3) The authors need to be careful about their claims without justification.
4) As you stated in your response, Lez spring discharge is a complex combination of natural discharge, pumping and legally regulated minimum discharge from the extracted water to protect downstream ecosystem. Therefore, the identification of the key input data for the ANN models is substantial.

We thank the Editor for this assessment. We do not answer to the points in detail, because they refer to referee comments, which are discussed and answered in detail in the following.

Response to Referee #1: by Kuo-Chin Hsu

The manuscript proposed to use convolutional neural network (CNN) associated with gridded meteorological data for Karst spring discharge modeling. CNN was applied to three karst spring watersheds in Europe. Results of 2D CNN model associated with gridded meteorological cells were compared to that of 1D CNN using climate station input data.

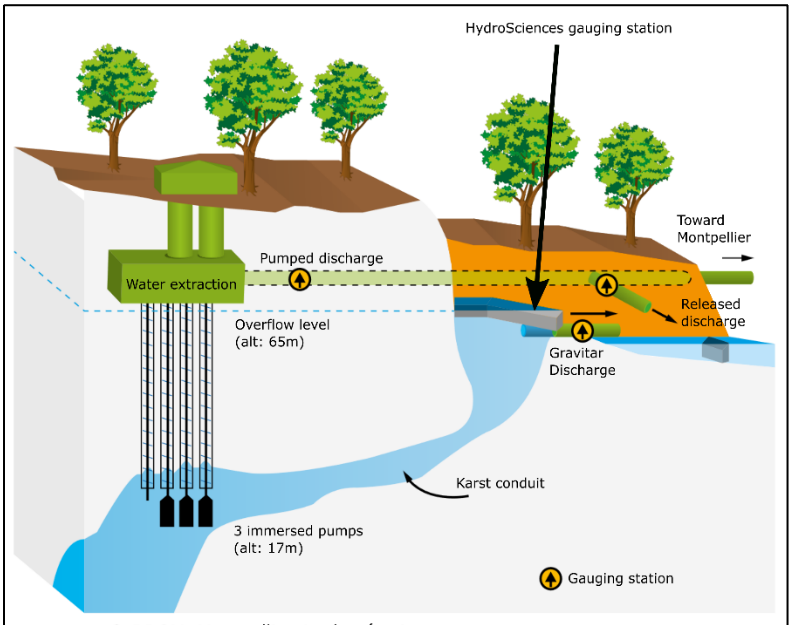
The manuscript is well written and technical sound.

General comments:

- CNN is a mature data-driven tool which highly relies on data availability and quality. The authors argue that less data is needed in the proposed approach to obtain satisficing results compared to previous deep learning approach and overcome the short of data from climate stations. The results show that 2D modeling is not necessary better than that of 1D and previous modeling in Lez spring. A question raised is that whether the key input data has been identified. For example, pumping may play an important driving factor but is not included in training and screened out by Bayesian model. Gridded meteorological data may not be enough to improve the model performance. The authors needs to address their contribution. Guide line for data preparation will be helpful for the suggestion of use machine learning.

Thank you for this thoughtful comment on data aspects. We think indeed that the 2D approach can overcome difficulties in climate station data availability. As we state in the manuscript: “climate stations are often not available within the catchment of a spring, do not match the data availability of the discharge time series (period or temporal resolution), or are more distant and thus do not truly represent the events in the catchment itself”. Nevertheless, we want to clarify that we do not think that the 2D approach needs less data, instead we think that rather the amount of work necessary to collect and preprocess the data is strongly reduced. Gridded meteorological data is available online and needs only minor preprocessing in contrary to most climate station data.

We agree that the results of the 2D approach are not necessarily better, as it can be seen in the example of Lez spring that you mention. It is true that Lez spring discharge is a complex combination of natural discharge, pumping and legally regulated minimum discharge from the extracted water to protect downstream ecosystems. Pumping is performed directly in the karst aquifer, thus also during dry periods with zero discharge, the legally regulated minimum discharge however is part of this extracted water. The legally regulated minimum discharge is released downstream into the Lez river (after the spring) and basically has no effect on the spring discharge other than being part of the pumped water volume. Please see the following illustration:



© BRGM, Montpellier Agglomération, *Edited by Cinkus (2020)*

Despite all these factors, we showed, that pumping is not a necessary input, to achieve highly accurate (NSE and KGE > 0.86 for the 1D model and > 0.75 for the 2D model) results for Lez spring and that we were able to simulate the discharge with solely meteorological data, using both the 1D and 2D models. We therefore think that focusing on climatic inputs allows a more consistent approach for all three study areas. Moreover, the focus of our study was to compare 1D meteorological inputs of climate stations to gridded 2D meteorological inputs, and not to achieve the best performance for each of the test sites. Another advantage of not including pumping data is that our approach can potentially be transferred on real forecasting tasks using weather/climate forecasts on different time scales in the future. For such applications, future pumping rates are not available. The most important reason not to include pumping is that the available data begins in 2011, which means we would loose three years of data in total for an already comparably short time series. We have now added a detailed explanation of not considering pumping as an input parameter to the manuscript. (L 160f, 407ff.)

In this specific case of Lez spring, a lot of work was necessary to produce the 1D precipitation input time series (compare Appendix B - Lez Catchment Precipitation Interpolation) due to very patchy climate station data. It seems that these additional efforts pay off in terms of higher performance compared to the 2D model. Nevertheless, the 2D climate data seems to offer a sufficient substitution, if needed.

We do not think that we can provide a general guideline for data preparation, because this step strongly depends on the datasets that a specific user intends to use. Nevertheless, we have included a description of the input data format that is used in our published python scripts, which should enable future users to adapt and apply them.

- The modeling uncertainty is quite low to almost without uncertainty that seems abnormal. The authors may explain this.

We apologize that, given our current formulation, this aspect does not become clear. The shown model uncertainty is derived from an ensemble of 10 differently initialized models, each using Monte-Carlo dropout to produce an ensemble of 100 different forecasts (so 1000 in total). For each of the 10 models we calculated the 95% confidence interval of the 100 available forecasts (1.96 times standard deviation, because of sigma rule for 95% confidence). What is shown in the final plots, is the 95% uncertainty of the mean of all 10 model ensembles, which is indeed very small. In the revised version of the manuscript, we have shortly clarified this aspect (Line 231ff.). We want to add that the shown uncertainty does not include other sources of uncertainty of the models (such as input data uncertainty). For this reason, it might seem abnormally small to you at first glance. We have also added a clarifying statement on this aspect in the revised version of the manuscript (Lines 300-302).

Specific comments:

- Line 203, write the long short term memory for abbreviation of

The LSTM as abbreviation is now introduced in the introduction. Thank you for pointing out. (L. 48-50)

- Lines 213-215. Although the all programs are available from Python community. Technical functions should be described for the used library or framework should be explained.

We are not quite sure what you mean exactly. We tried to explain the most important technical functions (like CNN) in the methodology section and give the associated references. Given the limited space for the manuscript, we avoided an in-depth explanation of all details and hope that referencing the used packages and frameworks in combination with the published Python code is sufficient. Additional necessary information should become clear to the python-experienced user from our published code.

- It is not clear functions for training, validation, optimization and testing periods in Table 1. The should be explained in main text.

Yes, you are right. Sorry that we missed to explain the purpose of those sets. We added a clarifying statement in Lines 242-245 in accordance with the comment on that aspect of Reviewer#2.

Response to Anonymous Referee #2

This study uses convolutional neural networks (both 2D and 1D) to model streamflow in three karst spring catchments. They compare 1D CNNs, which process time series of weather station forcing data, to sequential 2D – 1D CNNs, which process gridded meteorological forcing data from climate reanalysis. The use of widely available gridded meteorological data is advantageous due to its more complete spatiotemporal availability, in comparison to data from weather stations.

This is an interesting study with potentially applicable results; however, there are several key limitations with its current form. The literature review at present is insufficient and does not provide enough information to explain the relevance or context of the results. In many cases, the likelihood of hypotheses and claims is stated without justification. Additionally, conclusions surrounding the potential of these models for karst catchment localization and delineation are currently overstepping the results shown.

Major comments

Overall, the introduction is very brief and does not provide adequate context for the current work. Johannet et al (1994) is referred to but it is not stated what they did. The authors can expand to consider the further history of ANNs in water related research (e.g. Hsu et al 1995, Maier and Dancy 1996, Zealand et al 1999, Maier and Dandy 2000 and the references therein), which can lead to the use of deeper networks in water research. Additionally, 1D CNNs have been used for streamflow prediction in the past by others who the authors do not refer to (e.g. Hussain et al 2020, Van et al 2020). By further fleshing out the relevant history the authors can more convincingly present the relevance and potential applications of their work. We admit that the introduction was very brief, since we specifically focused on the application of CNN on karst spring discharge modeling. We now have added a statement on the study of Johannet et al. (1994) and extended the literature review on ANN application in water resources related research. We hope that this now better fulfills the requirements of presenting the relevance of our work more convincingly. Further we now provide references of 1D-CNNs in groundwater and streamflow/runoff modeling (Lines 22f., L25ff).

Why do the authors use a 1D CNN to learn temporal features rather than an LSTM, despite the many successes of LSTM-based modelling for streamflow prediction (e.g. Kratzert et al. 2018, 2019a, 2019b; Gauch et al. 2021, Frame et al. 2021, Anderson and Radic 2021; note that this list is not comprehensive or that all required but these papers can be a starting point for the authors to consider)? While there is a brief mention that 1D-CNNs are fast/stable, there is little mention or consideration given to the vast success of the LSTM approach at both daily and hourly scales, which is surprising given their prevalence. This study uses only three catchments and so I can't imagine that the training time between a 2D – 1D CNN model vs a

CNN-LSTM model would be prohibitively different. There is opportunity here to compare the two approaches quantitatively to see which performs better (e.g. perhaps a CNN-LSTM model will be able to simulate the streamflow peaks around Oct 2020 in Figure 3), or if there are differences in the "catchment delineation" results. While the authors don't absolutely need to perform this comparison, it would certainly help to justify their methodological choices (2D-1D CNN vs CNN-LSTM).

Thank you for this comment, which is completely understandable, especially given the successes of LSTM models, some of which you mentioned yourself. In preliminary work we did indeed also test LSTM models in combination with 2DCNNs and you are correct, for the 2D-models there is no large difference in training time (at least for these three sites) between 2DCNN+1DCNN and 2DCNN+LSTM. However, we also did not find systematic superiority or performance differences between LSTM and 1DCNNs as subsequent models to the 2DCNNs. Moreover, we also have shown in the closely related application of groundwater level forecasting, that 1D CNNs have in most cases an equal performance to LSTMs (Wunsch et al., 2021). Van et al (2020) also show the potential of CNNs compared to LSTMs in case of rainfall runoff modeling.

To be methodologically consistent, we therefore decided to neglect 2DCNN+LSTM models. This way, we do not change the forecasting model itself, but only replace the climate station input data by a 2D-CNN model, which learns relevant data itself. The primary goal of our study was to show that spatially distributed climate data can act as a reliable substitute in case of bad availability of climate station data. By not mixing up different model types, in our opinion a comparison between 1D and 2D-based models is more convincing. Though it would of course be possible to compare the CNN vs. the LSTM approach quantitatively to see which performs better, we would like to avoid this, as it would be just a different kind of research aspect and we prefer to keep the focus and the comparison of 1D vs. 2D meteorological inputs.

We have extended the justification of model choices in Lines 219-230.

References:

Van, S. P., Le, H. M., Thanh, D. V., Dang, T. D., Loc, H. H., and Anh, D. T.: Deep learning convolutional neural network in rainfall–runoff modelling, Journal of Hydroinformatics, 22, 541–561, https://doi.org/10/ggskkh, 2020.

Wunsch, A., Liesch, T., and Broda, S.: Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX), 25, 1671–1687, https://doi.org/10.5194/hess-25-1671-2021, 2021.

In many cases, the authors assert the likelihood of a claim without justification. These instances either need further exploration or to be reframed as "possible" rather than "probable". For example, in line 297 it is stated that the main source of uncertainty is "probably the uncertainty of parameter values resulting from the ERA5-land grid cell sizes". How are the authors confident that this is "probable"? Are the authors referring to the uncertainty, or error? There are other places where the authors describe a result or hypothesis as being "probable" without any justification. These instances are more conjecture than a discussion of probability (e.g. line 344, line 359, line 363, among others) and are not very convincing without additional support. Furthermore, there are instances where highly certain language is used without quantification (e.g. "perfectly" in line 297) or a contradictory mixture of language is used (e.g. "probably perfectly" in line 360; "quite exactly" in line 411). These should be changed as well. Another example is in line 407, where the authors suggest that the shape of the sensitivity pattern may be a "relic from spatial correlation of precipitation events". Can this hypothesis be supported or explored with evidence? The authors have the needed precipitation data so I am not sure why this conjecture is here without support.

Thank you for pointing out. We indeed used inaccurate wording at several passages, which we have revised throughout the complete manuscript. Please see the track changes document for a convenient overview of the changed wording.

See for example Lines: 290f., 300f., 309f., 372f., 375f., 391f., 395f., 421f., and others

Please also find our statement on the shape of the sensitivity pattern at Aubach spring in Lines 432-440.

I have concerns that the sensitivity methods applied do not actually work to "delineate" or "identify" the catchments to the degree that the authors are claiming. One key difference between Anderson and Radic (2021) and this study is that the basins in Anderson and Radic (2021) are in different regions of the input space, while the basins here are all centered (Figure 6). Having stations in different regions of the input was key to interpret if the model was learning to focus on the right regions in Anderson and Radic (2021); e.g. in Figure 7 in Anderson and Radic (2021), the model is sensitive in different regions which tells us that the model is learning different things for different stations. In Figures 6 and C1, one could argue that the models here have learned to generally focus on the central area under all circumstances and it is just coincidence that the basins are centered there as well. As I see it, in order to make steps towards catchment delineation, the authors need to demonstrate that the model automatically focuses on (1) the right location and (2) have the right "area of sensitivity". Neither of these points are quantified in this work, while both are referred to qualitatively; however, both can (and should) be more rigorously investigated. To point (1), the authors could run different tests with the basins placed at different locations in the input (e.g. 9 (or more) models could be made for each basin – one with the basin located in the top left area of the input, one with the basin located in the top center area of the input, etc). Then, the location of maximum sensitivity can be quantified and compared to the location of the basin. In this way (or in some similar way), the authors could more concretely conclude whether their approach is learning to focus on the correct area of the input or not. To point (2), the authors could run different tests with different input areas (e.g. for each basin, the authors could double/triple/etc the number of pixels in the x- and y- directions). Then, the "most sensitive area" can be quantified (e.g. the area greater than the half-maximum sensitivity, or some alternative metric) and compared with the basin area. The authors can find: is the most sensitive area always comparable to the known basin area? By addressing these two points (either as described above or in some other way), the authors can more convincingly state whether the CNN approach has potential for catchment delineation.

We understand your concerns that the model might just learn that the centering region is important. We followed your advice and performed additional analyses on all models, except the fine-resolution model for Aubach spring. This would have exceeded our computational possibilities for the short time available for revising this manuscript. We decided to focus exemplarily on the results of Unica spring in the manuscript, which nicely demonstrate that our models can indeed learn the correct location, regardless of the position of the catchment (Figure 7, Lines 458-478). For the sake of completeness, we added selected results for Lez spring and Aubach spring to the appendix (Figures C2 and C3). We now explain, what potential our approach bears for catchment delineation but also more clearly communicate that in its current form it is only useful for roughly localizing the position of the catchment (Lines 12, 67-68, title of section 3.4, 281-282, 455-456, Figure 7

Regarding point (2) of your comment: We think that this point is hard to evaluate given the results obtain in this study. A threshold (such as the half-maximum sensitivity you proposed) of the "most sensitive area" would hardly be transferable a) between parameters (as it strongly depends on the autocorrelation of different inputs, e.g. T is usually more auto-correlated than P) and b) between specific regions (in mountainous regions, meteorological parameters are less auto-correlated than in flat landscapes). Though this is an interesting idea, and we will keep in mind for future research, we do not think it is possible to include it in the current study. We rather think that a follow-on study with more catchments of adequate size should be performed, with a special focus on the delineation strategy development (Lines 487-490).

Additional comments:

Paragraph starting at line 43: This paragraph can be hard to follow when very little has been done to describe the architectures (e.g. the acronym 'LSTM' has not been defined).

Thank you, we now introduced the LSTM acronym in the introduction and rephrased this paragraph, to improve its understandability. L 47-57

Line 55: The authors state ANNs to be superior for points i) and iii), but no justification is given as to why they expect this.

Indeed, our formulation is not well justified and therefore misleading. In fact, we think that the most important advantage of our ANN-approach over a pure event-correlation is that ANNs are able to represent non-linear relationships. We rephrased this sentence. L 56-60

Line 200: Add references for batch normalization (e.g. Loffe et al, 2015) and dropout (e.g. Srivastava et al, 2014)

Thank you. We have added these references (L. 216, 217)

Line 203: LSTMs are claimed to be slower and with similar performance as compared to 1D CNNs for "this specific application". Does that mean that the authors have used LSTMs for streamflow modelling in karst catchments as well?

As mentioned above, regarding the extend justification of choice of methods, we have rephrased this paragraph. We hope it becomes clearer now, which experience and studies we draw our conclusions from. (L 219-228)

Section 3.2: It would be very useful to have an overview of the models that were used. Currently in Table 1 there is the time splitting scheme. In addition, it should be more clearly listed the length of the input sequence, number of observations, and number of parameters in each model (or layer).

Thank you for pointing out this important aspect. We added the number of observations to Table 1. Hyperparameters and information on the number of model parameters is now available in the appendix (Table D1).

Section 3.4: This section is very brief and does provide much context for the methods chosen (e.g. why follow Anderson and Radic, and not other interpretability methods?). Some statements are vague (e.g. "In short it works by perturbing spatial fractions of the input data by using a 2D-Gaussian curve" – what is meant by 'using'?). The final few sentences are written with certainty, although the methods have not been applied yet (e.g. "… a smaller area will most certainly have a higher influence on the spring discharge…"). This section can be challenging to follow, and I suspect especially so for readers who are not as familiar with Anderson and Radic (2021).

We apologize for not providing sufficient context. We have improved the section, added justification for following the approach of Anderson and Radic (2021), clarified the wording ("using") and relativized the certainty of our last statements. We also better explain why we modified their sensitivity approach to perturb only single input channels at a time. We hope the whole section is now easier to follow and more precise. L 268-282

Line 254: How is “satisfying fit” qualified? Satisfying as compared to what?

Thank you for pointing out. We removed this inaccurate wording and rephrased the section. Furthermore, we revised the complete manuscript and replaced inaccurate wording.
L 286. We also changed e.g. L328ff.

Line 276: It is not surprising that the models are within the same “order of magnitude”.
An order of magnitude of NSE spans 0.1 through 1, which is a huge range of performance.

We apologize for that mistake, which comes from inadequate translation of a figure of speech. We meant “in the same range”. We corrected the wording accordingly. L 310

Sections 4.1 – 4.3: These sections are a mixture of results and discussion. While that is not inherently an issue, both results and discussion are mixed throughout each section in ways that vary from section to section (e.g. 4.1 begins with results, but 4.2 begins with discussion before even a description of Figure 4). It would be easier to read and follow if the authors were more consistent between sections (e.g. first have results of 4.1, then discussion of 4.1, then results of 4.2 etc).

Thank you for pointing out. We have restructured this part, to provide a consistent presentation of results and discussion. Please see Section 4.2 in the tracked-changes document.

Line 331: It seems the authors mean “substantially” and not “significantly” as there is no discussion of statistical significance here. “Significant” is also referred to in author places where the authors do not appear to mean it in the formal sense.

Thank you for pointing out. We replaced several inaccurate usages of “significant” throughout the manuscript. Please see the tracked-changes document for all changes and for example the following lines: 122, 177, 335, 348, 354, 367, 385, 407, 494, 504

Line 339: If this new model is not going to be discussed or explored clearly, then it should not be brought up at all.

We agree, thank you. We removed this part from the revised version.

Lines 371 – 377: This is description of other studies should be moved to the introduction.

We admit that this is not the best place for describing the studies. We moved part of this description to the study area section, but also mention these studies now in the introduction (Lines 162ff., 62ff.). We hope that you agree that it is now appropriate.

Section 4.4 (and Section 3.4): It is not clear how the authors are defining catchment delineation/identification. Are they referring to the areas of the sensitivity fields that are greater than the half-maximum of sensitivity? For which input variable? It should be clarified just how the step from sensitivity heat map to catchment delineation could be made.

Currently we do not quantitively evaluate the sensitivity heatmaps, but judge only visually, if the known catchment coincides well with the high-sensitive areas on the heatmaps. So maybe "identifying the approximate catchment location" is a better formulation than "catchment delineation" in the present state. We generally adapted the wording e.g. in Lines 54-46, Renamed section: "Spatial Input Sensitivity and Catchment Localization", 68, 264, 281f, 456

Primary purpose of this study was to show the usefulness of the 2D approach for discharge forecasting. Identifying the catchment location was a very interesting additional aspect, which seems to be worth investigating further in the future. We clarified some thoughts on how to move from our results to a delineation strategy (L473-490)

References

Frame, J., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall-runoff predictions of extreme events, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2021-423, in review, 2021.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-25-2045-2021, 2021.

Hsu, K., Gupta, H. V. and Sorooshian, S.: Artificial Neural Network Modeling of the RainfallRunoff Process, Water Resour. Res., 31(10), 2517–2530, doi:https://doi.org/10.1029/95WR01955, 1995.

Hussain, D., Hussain, T., Khan, A. A., Naqvi, S. A. A. and Jamil, A.: A deep learning approach for hydrological time-series prediction: A case study of Gilgit river basin, Earth Sci. Informatics, 13(3), 915–927, doi:10.1007/s12145-020-00477-2, 2020.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22(11), 6005–6022, doi:10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S. and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resour. Res., 55(12), 11344–11354, doi:10.1029/2019WR026065, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23(12), 5089–5110, doi:http://dx.doi.org/10.5194/hess-23-5089-2019, 2019b.

Loffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv, 2015.

Maier, H. R. and Dandy, G. C.: The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters, Water Resour. Res., 32(4), 1013–1022, doi:10.1029/96WR03529, 1996.

Maier, H. R. and Dandy, G. C.: Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, Environ. Model. Softw., 15(1), 101–124, doi:https://doi.org/10.1016/S1364-8152(99)00007-9, 2000.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res., 15(1), 1929–1958, 2014.

Van, S. P., Le, H. M., Thanh, D. V., Dang, T. D., Loc, H. H. and Anh, D. T.: Deep learning convolutional neural network in rainfall–runoff modelling, J. Hydroinformatics, 22(3), 541–561, doi:10.2166/hydro.2020.095, 2020.

Zealand, C. M., Burn, D. H. and Simonovic, S. P.: Short term streamflow forecasting using artificial neural networks, J. Hydrol., 214(1), 32–48, doi:https://doi.org/10.1016/S0022-1694(98)00242-X, 1999