Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-113-RC2, 2021

Comment on hess-2021-113

Anonymous Referee #2

Referee comment on "Evaluation and interpretation of convolutional long short-term memory networks for regional hydrological modelling" by Sam Anderson and Valentina Radić, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-113-RC2, 2021

General comments:

This is an intriguing study that combines two distinct deep-learning technologies (the convolutional neural network, CNN, and long short-term memory neural network, LSTM) to create a new method for regional daily streamflow prediction that integrates complex spatiotemporal structures and dependencies. The method is applied to streamflow data from the southern portion of Canada's two westernmost provinces, which is a geophysically complex and interesting region. Some effort is also made to address physical interpretation and meaningfulness of the technique. It is a promising study with widely relevant results that has strong potential for publication in a top-tier hydrology journal like HESS.

That said, the submission as it currently stands appears to have some substantial issues that need to be addressed before it can be considered for publication. The overall feel of how the manuscript is written is one of technical naivete and oversimplification, undermining the credibility of the study. For example, the text of the paper and possibly some of the analytical steps suggest a superficial understanding of the physical hydrology of western Canadian rivers and their associated datasets; and overall, the literature review around machine learning and its hydrologic applications is wholly inadequate and does not provide the reader with accurate and meaningful context to the study. Additionally, several basic elements one normally expects of a machine learning paper today seem to be missing, like clear descriptions of training vs. testing vs. validation data subsets, or the use of informative benchmark models to evaluate the new model against. The study is also not reproducible based on the limited information provided in the paper.

My recommendation is to accept the paper for publication in HESS pending major revisions. I hope the detailed comments provided below, as well as the references section that follows those detailed comments, will be helpful to the authors as they revise their manuscript.

Detailed comments:

* Line 30: Should also cite Hsu et al. (1995) here, as to my knowledge it was the first peer-reviewed journal paper to present the use of machine learning for rainfall-runoff modeling. (Full literature citations are provided below.)

* Lines 34-37: This feels like an overstatement/misstatement of both the limitations of conventional machine learning and the advantages of deep learning in a hydrologic prediction context. For one thing, a basic result in AI, dating back to the late 1980s or so, is that non-deep ANNs (in particular, multilayer perceptrons having a single hidden layer) are theoretically capable of learning any continuous relationship. Another issue: contrary to what is implied in the passage, non-deep ANNs are not the only kind of non-deep machine learning – there are several other major classes (random forests, support vector machines, and so forth). There also continues to be intense research in non-deep ML to create new kinds of AI, including news kinds of neural networks, having certain useful characteristics that have been successfully applied to river prediction; online sequential learning is an obvious example (e.g., Lima et al., 2015, 2016, 2017). Indeed, new kinds of non-deep machine learning algorithms are being developed specifically for hydrometeorological analysis and prediction tasks (e.g., Cannon, 2010, 2011, 2018; Fleming et al., 2015, 2019, 2021). On the other hand, deep learning applications in hydrology are currently in vogue and seem to be very promising in certain circumstances, but the body of work on the subject – particularly around streamflow prediction – remains exceedingly small, and the ultimate suitability of deep learning to this task, including capabilities and limitations, remains unclear at this point. A more mature way of looking at deep learning in hydrologic prediction is that work to date suggests it is a promising research direction that could potentially offer an alternative or complementary approach to non-deep machine learning for certain tasks.

* Lines 49-50: The use of point observations (of weather, presumably) does not necessarily imply that a model is spatially lumped. It is very common in process-based hydrologic modeling, including semi-distributed and fully distributed models, to spatially interpolate measurements from point data sources. In fact, some process-based models even integrate that spatial interpolation step into the software platform, along with adjustments for adiabatic lapse rates, etc., etc.

* Lines 67-70: Explainability is an issue for all machine learning models, not just deep learning models; it feels like this passage is conflating ML generally with DL specifically. For a recent example of a new non-deep ML technique specifically introduced to improve interpretability of a practical hydrologic prediction model, see Fleming et al. (2021), which also provides a much better explanation of exactly why geophysical explainability is a key requirement for practical applications of machine learning in hydrologic prediction.

* Lines 78-79: the authors are not using the terms white-box and (in particular) black-box in the way they are usually used. Most working in hydrology, in particular, would regard

any physically explainable ML as being white-box in some sense. The term “black-box” is normally reserved for machine learning algorithms that do not offer any physical interpretability, which is to say, most of them.

* Lines 85-86: would be useful to note the similarities and differences between recurrent and LSTM neural networks here for a general readership. The text seems to be haphazardly switching between the two, which are related but not identical; LSTM is essentially a specific and advanced form of recurrent ANN. This applies to the title of the paper too; why "recurrent" instead of "long short-term memory"?

* Lines 104-105 are a bit off as well. There seems to be an implication here that more complex models are better models, and that in contrast this study is aiming for parsimonious models. That’s an odd way of looking at the desirability of different modeling approaches and structures. Most modelers view a parsimonious model as being fundamentally better, holding all else equal, i.e., so-called Accom’s razor.

* In addition to the various other papers referenced in this review that should be cited in the paper but were not, the authors may also wish to read and cite the review articles by Reichstein et al. (2019) and McGovern et al. (2019). Citing prior applications of machine learning to hydrologic and related modeling in the study area would also be appropriate. Some examples that come to mind include Rasouli et al. (2012), Lima et al. (2015, 2016, 2017), Snauffer et al. (2018), Fleming et al. (2015), Hsieh et al. (2003), and Shrestha et al (2021).

* Figure 1 would be much better, especially for an international readership that is unlikely to be strongly familiar with the study area, if it was a multi-panel figure that additionally illustrated topography, mean annual temperature, mean annual precipitation, and perhaps mean April 1 snow water equivalent.

* Lines 137-139: perhaps this passage merely is poorly written, but as it stands, the text implies a disturbing lack of understanding of the streamflow data being modeled. Naturalized flow data are flow data that have been adjusted for upstream water management activities – diversions, withdrawals, reservoir operations, etc. Data for stations upstream of dams are not necessarily naturalized, contrary to what is implied in this passage of the paper, and certainly in datasets like the HYDAT database used here, that step has not been undertaken and in many cases is unnecessary. Similarly, dams are not the only disturbance that result in non-natural streamflow data that would in principle require naturalization prior to use in a hydrologic modeling study of the sort done here; another obvious example is land use change. Why not use the Reference Hydrometric Basin Network (RHBN) stations or something similar? There is no mention here at all of the RHBN station network, which has been very widely used for decades for hydrological analysis and modeling studies in Canada. Also, I think quite a few hydrologists would raise their eyebrows at the specific data selection and processing procedures described in the first paragraph of section 3.1.

* The second paragraph of section 3.1 is also muddled. All that's needed here is a concise statement that hydrometric network density is much higher in southern than northern Canada, and so, for the purposes of this study, the authors focused on the former.

* While the approach described on lines 159-170 is interesting and perhaps sufficient for the purposes of this study, overall it appears to be a naïve representation of spatiotemporal pattern formation in streamflow regimes in this study area. At an absolute minimum, some acknowledgement of prior work, and some caveats about the simple method and assumptions used here for regime classification, are needed. See in particular Halverson and Fleming (2015) and references cited therein. A particularly notable omission is that glacier-fed rivers are not identified as a distinct regime, whereas glacial cover is well-known to be a major control of streamflow dynamics in several areas within this region; see Moore et al. (2009), Fleming et al. (2016), Jost et al. (2012), and Bidlack et al. (2021).

* Section 3.1: I think reproducibility requires that the hydrometric station list used here be shown to readers. A table in an appendix or supplementary materials would be fine.

* Section 3.5: provide information about the latency of the ERA5 reanalysis product – is it available in near-real time? Some reanalysis products are, and some aren't. It's a crucial question if one were interested in operationalizing a hydrologic prediction system like this for actual use in flood forecasting or another similar practical hydrologic prediction application. If ERA5 products are not available in near-real time, then briefly but clearly state that limitation and its implications for wider use of the modeling framework introduced here.

* "data" = plural

* Somewhere in Section 3 or 4 there needs to be an explicit and clear description of what the training vs. testing vs. validation datasets are. There is a very brief mention of training vs validation but it is inadequate. The reader is not provided with information about how the training vs validation split is made, nor whether another subset is reserved for out-of-sample hyperparameter selection. These are standard practices in machine learning, and information about them is needed for transparency, reproducibility, and credibility of the study.

* A modern paper on machine learning applications to hydrologic prediction requires, in general, a performance comparison against some relevant benchmark model. Linear regression using precisely the same input dataset as the deep learning method introduced here is an obvious starting point and can provide a meaningful assessment of how much nonlinearity, interactions, etc contribute to the (presumably better) performance of the new technique. A conventional ANN and an LSTM would also be useful, if more ambitious, points of comparison. The only significant attempt the paper makes at this is Table 2, which scours the peer-reviewed journal literature for examples of hydrologic models that

have been developed previously for a few of the locations considered in this study. That comparison is interesting and probably worth including in the paper, but it also has limited meaningfulness as different date ranges etc were used in the studies. Moreover, Table 2 relies on a small handful of academic studies and misses a lot of existing models within the study area operated by pragmatic water-management organizations like a large government-owned hydroelectric utility (BC Hydro), a provincial ministry (BC River Forecast Center), regional water management authorities (e.g., the MIKE-SHE model operated in the Okanagan Basin), and so forth. Moreover, given that even the simplest machine learning architecture outperforms process-based models in most cases, the somewhat mixed results in Table 2 are a little surprising. In Section 5 there is also a very brief verbal comparison against the LSTM-based work of Kratzert et al. (2018) but that study used a completely different set of basins and data, so again, the comparison is extremely approximate. I get that the purpose of this study is more around demonstrating a new technology, and perhaps delving a little into the question of explainability, but I suspect most readers would like to see more meaningful inter-model performance comparisons here.

* Estimating predictive uncertainty is a key element of a hydrologic prediction system. Figure 6 and its caption suggests that predictive uncertainty is quantitatively estimated here but is vague about the method. It appears that an ensemble of 10 different models is formed, and twice the standard deviation of the predictions from those 10 models on a given day is used as the de facto prediction bound for that day. This is a reasonable first-cut approach, I think. However, the method needs to be described in the methods section, and some capabilities and limitations need to be mentioned; I suspect that because weather uncertainty is not factored in (as far as I can tell from the manuscript as submitted) the ensemble spread will be substantially under-dispersive.

* The bar for explainability does not seem to be set very high here. The sensitivity analyses included in the paper are very useful, but they really amount to more of a plausibility test than an interpretability test. In particular, the paper demonstrates, though observing the CNN-LSTM responses to perturbations in the meteorological driving data, that its streamflow predictions (a) are most sensitive to weather in and near the basin as opposed to further away, and (b) are sensitive to temperature regimes, in particular, demonstrate hydrograph timing shifts corresponding to changes in snow accumulation and melt driven by temperature perturbations. Those results suggest the CNN-LSTM model is capturing key geophysical processes more-or-less correctly, but it does not clearly reveal physical explanations of the input-output relationships – only that the behaviors are consistent with some basic physical expectations. I think the paper is publishable without diving further into explainability, but the authors ought to phrase their outcomes a little more precisely around the question of interpretability and may wish to consider some additional sleuthing to demonstrate that the CNN-LSTM reveals physical processes. There is some precedent for this in machine learing-based streamflow modeling, and looking closely at those precedents may be useful to the authors; examples include Fleming (2007), Kratzert et al. (2018), and Fleming et al. (2021). Looking even more broadly across the literature than this would likely lead to even more suggestions of how to examine the geophysical relationships the model is capturing.

* Lines 629, "it is notable that the CNN-LSTM model achieves good streamflow simulation with only temperature and precipitation forcing data" – well, in practice the most widely applied hydrologic models tend to use only these two types of forcing because that's all

that is usually available, so I guess this point might be worth mentioning here but it’s not particularly “notable” to most streamflow modelers.

* Lines 635-638: is it possible that, through its empirical and complex meteorological input-hydrologic output mappings – effectively, a transfer function linking the meteorological data to the point streamflow observations – the CNN-LSTM effectively downscaled the reanalysis data, at least to some degree? May be worth talking about here.

* Lines 646-653: are the authors sure their method requires less data than an LSTM, as claimed here? Doesn’t the CNN-LSTM still ultimately need data for all N basins? This passage needs further explanation/clarification.

References:

Bidlack AL, Bisbing SM, Buma BJ, Diefenderfer HL, Fellman JB, Floyd WC, Giesbrecht I, Lally A, Lertzman KP, Perakis SS, Butman DE, D’Amore DV, Fleming SW, Hood EW, Hunt BPV, Kiffney PM, McNicol G, Menounos B, Tank SE. 2021. Climate-mediated changes to linked terrestrial and marine ecosystems across the Northeast Pacific Coastal Temperature Rainforest margin. Bioscience, doi.org/10.1093/biosci/biaa171.

Cannon AJ. 2010. A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. Hydrological Processes, 24, 673-685.

Cannon AJ. 2011. Quantile regression neural networks: implementation in R and application to precipitation downscaling. Computers and Geosciences, 37, 1277-1274.

Cannon AJ. 2018. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. Stochastic Environmental Research and Risk Assessment, 32, 3207-3225

Fleming SW. 2007. Artificial neural network forecasting of nonlinear Markov processes. Canadian Journal of Physics, 85, 279-294.

Fleming SW, Bourdin DR, Campbell D, Stull RB, Gardner T. 2015. Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific Northwest river. Journal of the American Water Resources Association, 51, 502-512.

Fleming SW, Goodbody AG. 2019. A machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water availability in the US West. IEEE Access, 7, 119943-119964.

Fleming SW, Hood E, Dahlke HE, O'Neel S. 2016. Seasonal flows of international British Columbia-Alaska rivers: the nonlinear influence of ocean-atmosphere circulation patterns. Advances in Water Resources, 87, 42-55.

Fleming SW, Vesselinov VV, Goodbody AG. 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.

Halverson MJ, Fleming SW. 2015. Complex network theory, streamflow, and hydrometric monitoring system design. Hydrology and Earth System Sciences, 19, 3301-3318.

Hsieh WW, Yuval, Li J; Shabbar A, Smith S. 2003. Seasonal prediction with error estimation of Columbia River streamflow in British Columbia. Journal of Water Resource Planning and Management, 129, 146-149.

Hsu K, Gupta HV, Sorooshian S. 1995. Artificial neural network modeling of the rainfall-runoff process. Water Resources Research, 31, 2517-2530.

Jost G, Moore RD, Menounos B, Wheate R. 2012. Quantifying the contribution of glacier runoff to streamflow in the upper Columbia River Basin, Canada. Hydrology and Earth System Sciences, 16, 849-860.

Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M. 2018. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences, 22, 6005-6022.

Lima AR, Cannon AJ, Hsieh WW. 2015. Nonlinear regression in environmental sciences using extreme learning machines: a comparative evaluation. Environmental Modelling and Software, 73, 175-188.

Lima AR, Cannon AJ, Hsieh WW. 2016. Forecasting daily streamflow using online sequential extreme learning machines. Journal of Hydrology, 537, 431-443.

Lima AR, Hsieh WW, Cannon AJ. 2017. Variable complexity online sequential extreme learning machine, with applications to streamflow prediction. Journal of Hydrology, 555, 983-994.

McGovern A, Lagerquist R, Gagne DJ II, Jergensen GE, Elmore KL, Homeyer CF, Smith T. 2019. Making the black box more transparent: understanding the physical implications of machine learning. Bulletin of the American Meteorological Society, November, 2175-2199.

Moore RD, Fleming SW, Menounos B, Wheate R, Fountain A, Stahl K, Holm K, Jakob M. 2009. Glacier change in western North America: influences on hydrology, geomorphic hazards, and water quality. Hydrological Processes, 23, 42-61.

Rasouli K, Hsieh WW, Cannon AJ. 2012. Daily streamflow forecasting by machine learning methods with weather and climate Inputs. Journal of Hydrology, 414/415, 284-293.

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. Nature, 566, 195-204.

Shrestha RR, Bonsal BR, Bonnyman JM, Cannon AJ, Najafi MR. 2021. Heterogeneous snowpack response and snow drought occurrence across river basins of northwestern North America under 1.0*C to 4.0*C global warming. Climatic Change, 164, 40.

Snauffer AM, Hsieh WW, Cannon AJ, Schnorbus MA. 2018. Improving gridded snow water equivalent in British Columbia, Canada: multi-source data fusion by neural network methods. The Cryosphere, 12, 891-905.