Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC1
https://doi.org/10.5194/hess-2021-176-AC1, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Reply on RC1

Farhang Forghanparast et al.

Author comment on "Deep, Wide, or Shallow? Artificial Neural Network Topologies for Predicting Intermittent Flows" by Farhang Forghanparast et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-176-AC1, 2021

Reviewer 1
Comment: The manuscript suggests the combination of classification and regression models (deep and wide topology) to increase the accuracy of the current data-driven models available for streamflow forecasting in intermittent rivers. Overall, the topic is very interesting, and the manuscript was written well. The suggested models are new, and the results are well discussed.

Response: We thank the reviewer for this comment and for finding our work to be interesting, innovative, and well-written. We also appreciate the reviewer's other detailed and helpful comments. We have made the necessary modifications as described below.

- **Comment:** line 98: update the references of the current regression-based models regarding the following paper: Mehr, A. D., & Gandomi, A. H. (2021). MSGP-LASSO: An improved multi-stage genetic programming model for streamflow prediction. Information Sciences, 561, 181-195.

Response: Thank you for your suggestion and pointing us to this relevant reference The references in Line 98 are updated and the mentioned reference is added. The paragraph at line 97 is updated as:

"Conventionally used models for intermittent streamflow forecasting only include the regression cell of the wide network (Cigizoglu, 2005; Kisi, 2009; Makwana and Tiwari, 2014; Rahmani-Rezaeieh et al., 2020; Mehr and Gandomi, 2021). This configuration typically has a single input layer, a hidden layer and an output layer and is referred to as shallow topology (or shallow model) in this study."

- **Comment:** line 118-119: In the hydrological modeling community ANNs are known as regressors; however, the authors claimed ANNs as high-performance classifiers. The given references in line 119 are out of the hydrological forecasting community. It is

better to remove lines 118-119.

Response: Thank you for your comment. We agree that, unlike regression, classification is not a common approach in hydrological modeling. We, therefore, provided citations from other fields where ANN classifiers have been used over a wide range of datasets in an effort to justify the testing of this approach in this application.

The paragraph at line 118 is re-written as:

"Unlike the regression approach, which is widely used for streamflow forecasting, classification is not a common methodology in hydrological modeling. Artificial Neural Networks (ANN) are however known to provide high performance in both regression and classification over a wide range of datasets and applications in other fields (e.g., Araulampalam and Bouzerdoum, 2003; Rocha et al., 2007; Landi et al., 2010; Al-Shayea, 2011; Amato et al., 2013; Wang et al., 2017; Bektas et al., 2017) and thus provide a strong basis for testing their use here. While the developed topologies in this study are independent of the algorithm, for the sake of brevity, the same family of ANN models was used for the regression and classification cells in this study."

- **Comments:** Furthermore, please justify why you don't select a well-known classifier such as SVM or random forest?

Response: Thank you for this important comment.

As we state in our paper, any classification and regression modeling scheme can be used with our proposed approach. As our focus was on the presentation of an integrated classification + regression methodology for modeling intermittent flows, a detailed comparison of suitable algorithms for classification and regression cells was outside the scope (but we plan to pursue this important question in a separate paper). Secondly, ANNs were chosen because they are known to perform well in both classification and regression tasks and picking a single approach helps maintain the brevity of the paper and keep the focus on the presentation of the coupling framework. We have added a comment to this regard in the manuscript.

The choice of ANNs was made here as they are known to perform both regression and classification tasks with a high degree of accuracy (e.g., Araulampalam and Bouzerdoum, 2003; Rocha et al., 2007; Landi et al., 2010; Al-Shayea, 2011; Amato et al., 2013; Wang et al., 2017; Bektas et al., 2017) and selecting a similar architecture helps with brevity and keep the focus of the presentation on the proposed modeling frameworks. However, the proposed approach is model agnostic and any other suitable classification and regression scheme can be used instead of the ANNs schemes used here for illustrative purposes.

- **Comment:** Section 3 is a part of the methodology of this paper. It could be combined with section 2.

Response: Thank you for your suggestion. We agree and the manuscript has been updated with "Parameter estimation for deep and wide artificial neural network architectures" as in section 2.6. All the following sections and

subsections are updated subsequently (Please refer to the attached PDF, the additional information for Comment 4, Reviewer 1).

- **Comment**: The authors must avoid providing literature review in this section and section 4 as well. For example, lines 203-209 must be removed, or lines 216-236 must be substantially shortened.

Response: Thank you for your comment. We have revised the manuscript, reduced some citations, and shortened the parts on "Greedy learning", "Extreme Learning Machine configuration", and "Regularization for robust estimation for hidden node selection" from lines 180 to 244. However, several choices can be made while training ANN architectures and we retained some references here to justify our choices and provide readers with suitable context and additional references to look at while replicating or extending this work.

The paragraphs from line 180 to 244 are revised as: (The parts in brackets are removed)

2.6.1.1 Greedy learning

Greedy learning is a widely used strategy in machine learning for training sequential models such as regression trees, random forests, and deep neural networks (Friedman, 2001; Hinton et al., 2006; [Bengio et al., 2006;] Larochelle et al., 2009; [Johnson and Zhang, 2013; Liu et al., 2017] Naghizadeh et al., 2021). In this approach, parameter estimation is not carried out on a global objective function but conducted in a piece-wise manner. This simplification reduces the number of parameters to be estimated and therefore makes the optimization problem mathematically tractable. [Despite the lack of a global objective function,] Greedy learning algorithms are known to produce useful machine learning models that exhibit a high degree of accuracy (Knoblock et al., 2003; Su et al., 2018; (Wu et al., 2018) Belilovsky et al., 2019).

Adopting the greedy learning approach here essentially decouples the global objective function (Eq. (9)) into two separate optimization problems whose objective functions are given by Eq. (7) and Eq. (8). In other words, the models in the classification and regression cells are fit separately to estimate the unknowns within each cell. Generally, the increased computation burden of solving two optimization problems is offset by the gains obtained by separating the overall search space of the global objective function. Therefore, the greedy optimization approach was adopted here to solve Eq. (9).

2.6.1.2 Extreme Learning Machine configuration

An Extreme Learning Machine (ELM) is a special form of MLP wherein the weights for the input-hidden nodes connections and the associated bias terms are randomly assigned, rather than being estimated via optimization. This strategy greatly reduces the complexity of the parameter estimation process as [the weights connecting the inputs to hidden nodes and the associated bias terms need not be estimated and] only the weights and bias associated with the output node need to be estimated.

From a conceptual standpoint, as the input-output computations (Eq. (2) and Eq. (3)) are not part of the parameter estimation process, they only need to be performed once. This is tantamount to applying a randomized nonlinear transformation to the original inputs to create a transformed set of variables (i.e., the outputs of the hidden nodes). As the hidden node-output sub-model is a logistic regression formulation in case of a

classification problem and linear regression formulation in case of a continuous output, the optimization can be performed with relative ease using analytical approaches.

[Despite the random nature of input-hidden node transformation, ELMs have been shown to have universal approximation capabilities (Huang et al., 2006; Cocco Mariani et al., 2019). From a practical standpoint, they are noted to perform well and provide results that are comparable to other machine learning methods, especially MLPs that have been fitted using nonlinear gradient descent approaches (Zeng et al., 2015; Yaseen et al., 2019; Adnan et al., 2019).]

ELMs are increasingly being used in hydrology for a wide range of problems (Deo and ¸ Sahin, 2015; Atiquzzaman and Kandasamy, 2015; Deo et al., 2016; Mouatadid and Adamowski, 2017; Seo et al., 2018; Afkhamifar and Sarraf, 2020), especially streamflow forecasting (Lima et al., 2016; Rezaie-Balf and Kisi, 2017; Yaseen et al., 2019; Niu et al., 2020).

The use of Greedy learning and ELM configuration greatly reduces the mathematical complexity of the parameter estimation process for the proposed deep and wide topologies for predicting intermittent flow time-series. However, the problem of overfitting (Uddameri, 2007) cannot be ruled out, especially when the hidden layer contains a large number of nodes. Overfitting must be addressed to ensure the proposed deep and wide topologies learn the insights in the training dataset and are able to generalize to other inputs that are presented to the model during the calibration phase.

2.6.1.3 Regularization for robust estimation for hidden node selection

[While the ELM greatly reduces the computational complexity, the randomization of input-hidden node weights implies that the overall model fits are subject to chance.] The number of hidden nodes is an important hyper-parameter that critically controls the performance of ANNs, in general, and ELMs, in particular (Huang and Chen, 2007; [Wrong et al., 2008;] Feng et al., 2009; Lan et al., 2010; [Zhang et al., 2012;] Ding et al., 2014). If the number of hidden nodes is set too low, then the improper specification of hidden node weights due to random selection is hard to correct. Having a large number of hidden nodes improves the chances of at least some of them having high weights. However, the nodes with the smaller weights tend to learn the noise in the data resulting in poor generalizing capabilities. Reducing overfitting while maintaining a sufficient number of hidden nodes to capture nonlinear input-output relationships using ELM has received a significant amount of attention in recent years (Yu et al., 2014; [Shukla et al., 2016; Feng et al., 2017;] Zhou et al., 2018; [Duan et al., 2018;] Lai et al., 2020).

[The second part of the ELM develops a linear least-squares relationship between the output of the hidden nodes and the ultimate output (predictand).] When there are a large number of hidden nodes, correlations between them are to be expected. The presence of correlated inputs results in multicollinearity issues when performing ordinary least squares regression (Hamilton,1992). Regularization approaches are commonly used to reduce the impacts of correlated inputs and have been used with ELMs to minimize the overfitting problem (Inaba et al., 2018; Zhang et al., 2020). In this approach, an additional term, which is a function of the weights connecting the hidden node and output weights, is added to the loss function (and is referred to as L-norm). The revised objective function (see Eq. (10)) not only minimizes the sum of squares of residuals but also the number of hidden nodes.

The L2-norm, also referred to as Ridge norm or Tikhanov regularization, is a function of squares of the weights (see Eq. (11)). This approach typically forces weights with small singular values to be small numbers (as close to zero as possible), which can be ignored during predictions. The L1-norm, also referred to as LASSO norm (Eq. (12)), minimizes

the absolute value of the weights and actually sets the insignificant weights to a value of zero. The loss function with L1-norm results in a convex optimization problem that can be solved via linear programming and, therefore, commonly adopted (Zhang and Xu, 2016). Furthermore, the L1-norm is shown to induce a greater degree of sparseness [than the L2-norm without sacrificing prediction accuracy (Fakhr et al.,) The L1-norm is also] and to be more robust to outliers in comparison to the L2-norm (Zhang and Luo, 2015). Outliers are of particular concern when dealing with highly variable intermittent flow. The value in Equation 10 is a weighting factor that denotes the relative importance of the regularization term vis-a-vis the error minimization term and can be obtained via cross-validation procedure (Martínez-Martínez et al., 2011).

- **Comment:** Regarding the organization of the manuscript, I prefer to see Figure A1, Table A1, and Table A2 within the main text. The manuscript does not need an appendix.

Response: Thank you for your comment. Based on your recommendation and in the interest of brevity, Figure A1, Table A1, and Table A2 have been moved to the supplementary material. The updated manuscript has no appendix. Please refer to the attached PDF, the additional information on Comment6, Reviewer1.

- **Comment:** Line 149: remove the full expression of ANNs as you already provided in line 118.

Response: Thank you for your comment. The full expression of ANNs is removed from line 118.

- **Comment:** Line 181-187: redundancy in the citation is seen in this paragraph. Remove some of them.

Response: Thank you for your comment. Some of the citations have been removed. The paragraph of line 181 is revised as:

"Greedy learning is a widely used strategy in machine learning for training sequential models such as regression trees, random forests, and deep neural networks (Friedman, 2001; Hinton et al., 2006; Larochelle et al., 2009; Naghizadeh et al., 2021). In this approach, parameter estimation is not carried out on a global objective function but conducted in a piece-wise manner. This simplification reduces the number of parameters to be estimated and therefore makes the optimization problem mathematically tractable. Greedy learning algorithms are known to produce useful machine learning models that exhibit a high degree of accuracy (Knoblock et al., 2003; Su et al., 2018; Belilovsky et al., 2019)."

- **Comment:** Remove capitalization of each word in section 4.4.

Response: Thank you for your comment. The capitalization of each word is removed from this title.

- **Comment:** Flow rate or flowrate? Use a fixed one in the whole text.

Response: Thank you for your comment. We have made modifications and used "flowrate" throughout the updated manuscript.

- **Comment:** In section 5, lines 341-342 are irrelevant. Please remove.

Response: Thank you for your comment. The sentence beginning in line 341, is revised as:

"Coarse-scale runoff estimates generated from an ensemble of regional Variable Infiltration Capacity (VIC) models were obtained and used as an input to condition model predictions (i.e., inform the model of the best initial guess of the likely streamflow)."

- **Comment:** At the end of Section 5, list the selected inputs clearly. Statistical features of inputs must be given.

Response: Thank you for your comment. These lines were added to the end of the "Input specification for deep and wide ANNs for predicting intermittent streamflows" part:

"Ultimately, precipitation, potential evapotranspiration, soil moisture index, and their lags as well as the VIC-estimated runoffs formed the final set of inputs used for each stream. Table S3 (in Supplementary Materials) provides a summary of these inputs and their statistical features at each station."

Please refer to the attached PDF, the additional information on Comment 11, Reviewer1, for TableS3.

- **Comment:** Section 7.1. Calibration must be replaced with training.

Response: Thank you for your comment. In the subtitle of Section 7.1. "calibration" is now replaced with "training".

Please also note the supplement to this comment:
https://hess.copernicus.org/preprints/hess-2021-176/hess-2021-176-AC1-supplement.pdf