Response to RC#3

[1] I found this manuscript is very confusing. I am not sure about their numerical experiments. Before I have a good understanding of their experiments I cannot give a good review on the results. Below are my comments for now. I am happy to give more detailed review after I have a better understanding of their numerical experiments from their revised manuscirpt.

Response: Thank you for your comments. We have answered your concerns point by point below, with the hope that you can better understand our intent.

[2] The title mentioned "quantifying uncertainties in geographic extrapolation". I am wondering how the authors quantified the uncertainties. This uncertainty quantification is one of the objectives of this study if I understand the authors correctly, but I did not see any related discussion in the introduction till the results analysis.

Response: Our aim was to vary input datasets used for forcing in order to quantify uncertainties in our ML algorithms in a geographic extrapolation topic. The ML algorithm had different hyper-parameters at the source region when we forced it with different input datasets. Variations in the hyper-parameters resulted in different streamflow predictions at the target based on the pre-trained ML algorithm at the source. The overall best ML algorithm in our experiment was obtained from the algorithm (which, in our experiment, was eXtreme Gradient Boosting) having the most accurate predicted range for the streamflow at the target region. We will enhance the introduction to highlight the need for uncertainty quantification in ML algorithms.

[3] The conclusion in the abstract said "This study provides insight into the selection of input datasets and ML algorithms with different sets of hyperparameters for a geographic streamflow extrapolation." I am wondering what the insights are specifically.

Response: There are two items discussed here. Firstly, we trained ML algorithms with different input permutations to determine the impact of input sets on the capability of ML algorithms in an entirely new prediction domain having an unknown input-output relationship. Secondly, since variation of input datasets can result in different hyper-parameters for ML algorithms, we examined how such variations in hyper-parameters impact the predicted streamflow range in the new study domain.

[4] The effectiveness of transfer learning depends on the similarity of the source and the target. I am wondering whether the authors performed a similarity analysis which I think it is important to analyze the effectiveness of the extrapolation. And it might explain that adding more sample data from the sources did not improve the performance in predicting the targets.

Figure S1 shows the spatial pattern analysis using Uniform Manifold Approximation and Projection (UMAP), which untangles the inputs for source and target regions for twelve different months. It is interesting to note that the target catchments (rectangles) are primarily within catchments from the source, demonstrating the possibility that pre-trained ML over the source regions (circles, crosses and plus marks) can predict the output at the target regions.

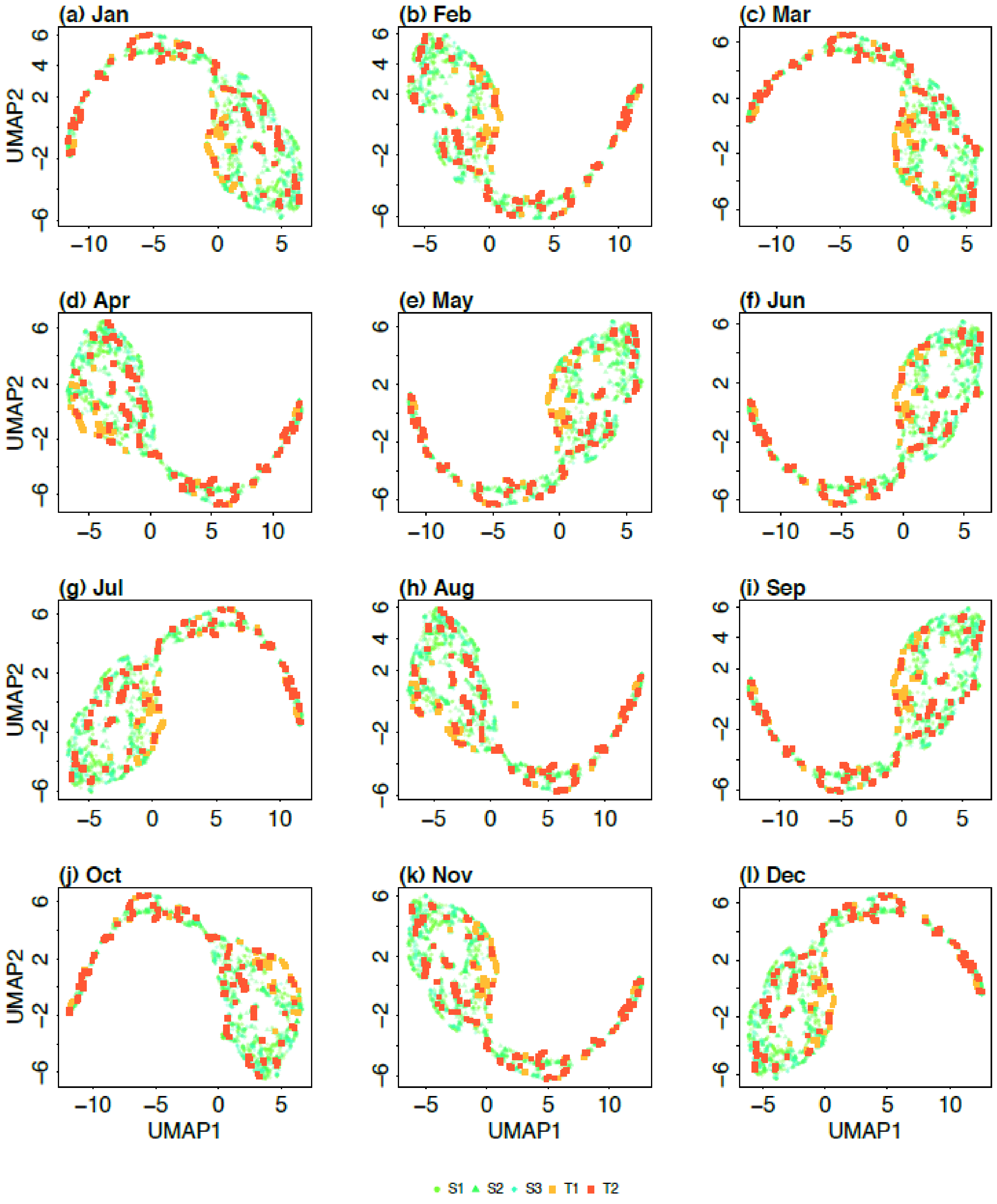

Figure S1. Uniform Manifold Approximation and Projection (UMAP) clustering of the source and target regions (colored and symbolized by catchment). Source Region 1 (S1)-North America, Source Region 2 (S2)-South America, Source Region 3 (S3)-Europe; Target Region 1 (T1)-South Africa, Target Region 2 (T2)-Central Asia

[5] Line 107, what “hypothesis”?

Response: The learned/pre-trained ML algorithm at the source may be able to predict streamflow in the entirely new study domain (the target).

[6] Why specifically chose these three ML methods? How about the more recently widely used LSTM network? It is known that these three chosen ML methods cannot learn the temporal dependence and the memory effects of the dynamic inputs on streamflow outputs.

Response: This study focuses on a spatial prediction model rather than a temporal model. That is why we do not see the fit of LSTM. In addition, our sample dataset is not too large to train LSTM. The sample size used to train the model in our experiment can be a relatively small number of samples (please see Figure 4), while LSTM may not perform well even with multiple decades of records (daily time step) (Kratzert et al., 2019).

Reference:

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12), 5089-5110.

[7] Did the authors consider the influence of lagged P and T on current streamflow when they designed the numerical simulations?

Response: We did not consider lagged time in our experiment. First, our experiment is not a time-series model but rather a spatial predictive model. In addition, we aggregated all variables to a climatological monthly time scale. With regard to results, lagged time analysis would not make sense.

[8] Please be specific about the input and output data. Both spatial and temporal data were considered, how the authors split the data for training-validation-testing in terms of both space (i.e., catchments) and time period. The description of 25%-25%-50% of the total number of data is very vague. I do not know what the total number of data represent?

Response: Each model has 16 input variables, including two dynamic variables (precipitation and air temperature) and 14 static/ invariant variables. Since our work was based on a climatological monthly timescale, we primarily focus on spatial scale but not temporal scale. For a typical experiment, the associated dataset was divided randomly by a ratio of 0.25, 0.25, and 0.5 for training, cross-validation and testing. We created 100 folds of these training, cross-validation, and testing datasets for each experiment. Table S1 below represents data characteristics for the first six input datasets (of 100 in total) from experiment 01 (EX01) and experiment 07 (EX07), respectively.

Table S1. Statistical description of input datasets for EX01 and EX07. The labels ‘train’, ‘cv’, and ‘test’ denote training, cross-validation, and testing datasets. The labels ‘n’, ‘min’, ’mean’, and ‘max’ denote total sample size, minimum, mean, and maximum values.

no	**train.n**	**train.min**	**train.mean**	**train.max**	**cv.n**	**cv.min**	**cv.mean**	**cv.max**	**test.n**	**test.min**	**test.mean**	**test.max**
EX01												
1	247	0.14	19.02	160.12	247	0.17	19.63	266.12	493	0.16	17.08	285.83
2	247	0.17	18.17	221.46	247	0.17	18.92	187.12	493	0.14	17.86	285.83
3	247	0.17	16.50	108.78	247	0.15	21.43	285.83	493	0.14	17.44	187.12
4	247	0.14	17.45	285.83	247	0.16	18.70	221.46	493	0.17	18.32	266.12
5	247	0.14	18.62	143.46	247	0.17	15.90	187.12	493	0.15	19.15	285.83
6	247	0.15	17.99	119.21	247	0.17	18.06	285.83	493	0.14	18.38	187.12
7	247	0.17	18.14	285.83	247	0.14	16.73	108.78	493	0.15	18.97	266.12
EX07												
1	564	0.14	34.90	372.5	564	0.16	38.61	337.41	1129	0.17	35.66	338.39
2	564	0.15	35.46	372.5	564	0.17	37.88	337.41	1129	0.14	35.74	316.24
3	564	0.16	34.66	372.5	564	0.23	34.87	301.52	1129	0.14	37.65	338.39
4	564	0.22	36.77	298.2	564	0.16	36.54	372.5	1129	0.14	35.75	338.39
5	564	0.16	35.13	372.5	564	0.15	34.57	337.41	1129	0.14	37.56	316.24
6	564	0.23	35.50	266.32	564	0.16	40.43	372.5	1129	0.14	34.45	337.41

[9] I am confused about the local-based models. It said “using target catchments to train the ML algorithms”, did it also include the source catchments or just target catchments?

Response: Local-based models only consider target catchments. We used these local-based models as benchmark results (our usual way of developing models) to examine the performance of the source-based models at the target.

[10] Figure 2. I am confused about the total data, i.e., training is about 25% of total. Did this total data include all five regions (source +target) or just source/target?

Response: The total data depend on the experiment; please refer to Table S1 in [8] for more detail. It should be noted that no experiment includes all five regions. This experiment was designed with the idea of developing pre-trained models from different source regions (North America, South America, and Western Europe). Uncertainties were analyzed after we used these pre-trained models to predict streamflow at the target region (Central Asia, South Africa).

[11] Table 3 and the 7 experiments need more explanation. I am not sure what these 7 experiments are.

Response: We designed different input imputations to get different pre-trained ML models (different hyper-parameters) at the source. Then we used these pre-trained ML models to predict climatological streamflow at the target. We will enhance our explanation for Table 3 in the revised manuscript.

[12] Line 241, for each of these 100 simulations, the hyperparameter tuning was performed and the best results were presented? Please clarify.

Response: Each simulation was forced by a different input set. Hyper-parameter tuning was performed and the metrics for the testing sets (unseen sets which existed during the hyper-parameter tuning process) were calculated. We will revise this sentence for clarification.