Corrigendum to Hydrol. Earth Syst. Sci., 26, 3377–3392, 2022
https://doi.org/10.5194/hess-26-3377-2022-corrigendum

Hydrology and Earth System Sciences
Open Access
EGU

Corrigendum to "Deep learning rainfall–runoff predictions of extreme events" published in Hydrol. Earth Syst. Sci., 26, 3377–3392, 2022

Jonathan M. Frame[1,2], Frederik Kratzert[3], Daniel Klotz[3], Martin Gauch[3], Guy Shalev[4], Oren Gilon[4], Logan M. Qualls[2], Hoshin V. Gupta[5], and Grey S. Nearing[6]

[1]National Water Center, National Oceanic and Atmospheric Administration, Tuscaloosa, AL, USA
[2]Department of Geological Sciences, University of Alabama, Tuscaloosa, AL, USA
[3]LIT AI Lab & Institute for Machine Learning, Johannes Kepler University, Linz, Austria
[4]Google Research, Tel Aviv, Israel
[5]Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, USA
[6]Google Research, Mountain View, CA, USA

Correspondence: Jonathan M. Frame (jmframe@crimson.ua.edu)

Published: 24 January 2023

1 Abstract

An error in the experiment setup code was found by the authors that led to an incorrect splitting of training and testing data. This error has been corrected, and changes from the original publication are explained here. This corrigendum maintains the same section names and numbering as the original publication but only includes text and figures that have been corrected.

2 Methods

2.1 Models

2.1.1 ML models and training

The original publication stated that the third training/test period split all training and test years in each basin by at least 1 year. We found an error in our code that created these 1-year splits; specifically, the Python notebook split_test_val_train.ipynb used the date 10 January (yyyy-01-10), rather than 1 October (yyyy-10-01), as the beginning of the annual period when splitting basin years into training, validation, and test sets. This error has been corrected in the code, and the affected figures are replotted in Sect. 3.1 and Appendix C in this corrigendum. It was also stated, in the original publication, that the third training/test period split used all water years in the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set with a 5-year or lower return period peak flow for training. This phrasing was erroneous. The third training/test period split used only the first 13 water years following 1981 with a 5-year or lower return period peak flow for training.

3 Results

3.1 Benchmarking peak flows

The correction in the data splits slightly alters the results presented in Fig. 2, which shows the average absolute percent bias of annual peak flows for water years with different return periods, from models with a training/test split based on return periods, with all test data coming from the water years 1996–2014. It is important to point out that the three models that are trained/calibrated on only the high-probability years (return periods less than or equal to 5) all show similar performance degradation as the event return period increases. This is consistent with the results when the training period was not split by the return period of peak annual flow, as shown in Fig. 1 of the original publication.

This error slightly alters the results presented in Figs. 2 and C2. The three models that were trained/calibrated using the data splits all had relatively similar changes to their per-

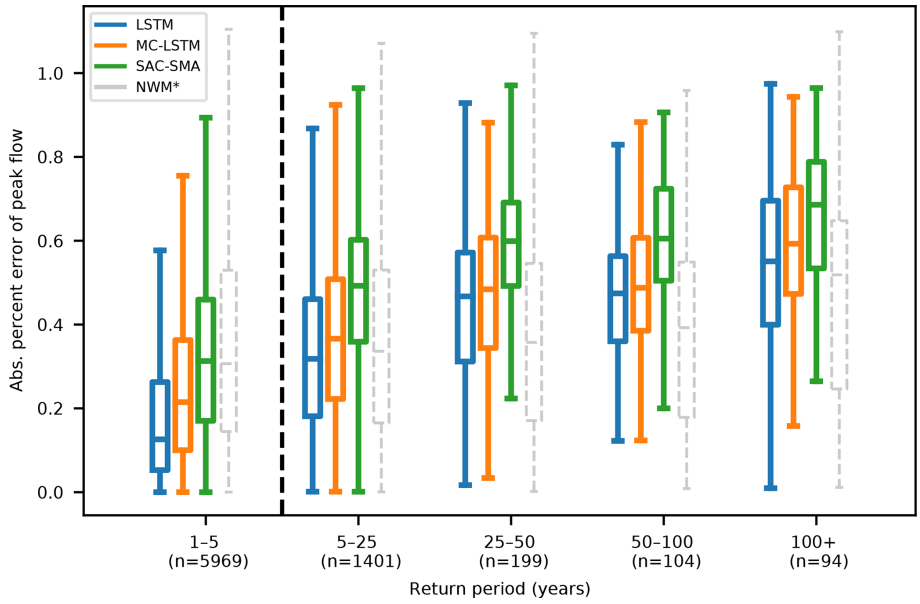

Figure 2. Average absolute percent bias of daily peak flow estimates from four models binned down by return period, showing results from models trained only on water years with return periods less than 5 years. The 1–5 year return period bin (left of the black dashed line) show statistics calculated on training data, while bins with return period years 5+ (to the right of the black dashed line) show statistics calculated on testing data. The LSTM, MC-LSTM, and SAC-SMA models were all trained (calibrated) on the same data and time period. The NWM was calibrated on with the same forcing data, but on a contiguous time period that does not exclude extreme events.

formance in the extreme-event scenarios. The US National Water Model CONUS Retrospective Dataset 2 (NWM-Rv2) was not calibrated specifically for this study, so those results do not change. In these corrected results, the NWM-Rv2 outperforms the long short-term memory (LSTM), mass-conserving LSTM (MC-LSTM) and Sacramento Soil Moisture Accounting (SAC-SMA) models in the lowest-probability events (return periods 25+); however, these results of the NWM-Rv2 and the other models presented are not directly comparable, as the NWM-Rv2 was calibrated.

Appendix C: Benchmarking annual return period metrics

Figure C2 shows the nine performance metrics calculated on model test results split into bins according to the return period of the peak annual flow event. The LSTM, MC-LSTM and SAC-SMA models were calibrated/trained on water years with a peak annual flow event that had a return period of less than 5 years (i.e., bin 1–5, indicated by the dashed line). The results shown in this figure are for the water years 1996–2014. The LSTM and MC-LSTM perform better than the SAC-SMA model according to every metric and for all bins. There are several instances where the NWM-Rv2 performs best, particularly for the lower-probability events. Although the NWM-Rv2 calibration does not correspond to the training/calibration period of SAC-SMA, LSTM or MC-LSTM, the NWM-Rv2 results are for reference and are not a formal component of the experiment.

Code and data availability. Interactive Python scripts that have corrected the aforementioned error, containing all post hoc analysis reported in this paper, including calculating metrics and generating tables and figures, are available at https://doi.org/10.5281/zenodo.7314083 (Frame, 2022).

References

Frame, J.: jmframe/mclstm_2021_extrapolate: Corrigendum for extreme events paper https://doi.org/10.5194/hess-26-3377-2022 (v1.0.2), Zenodo [data set], https://doi.org/10.5281/zenodo.7314083, 2022.

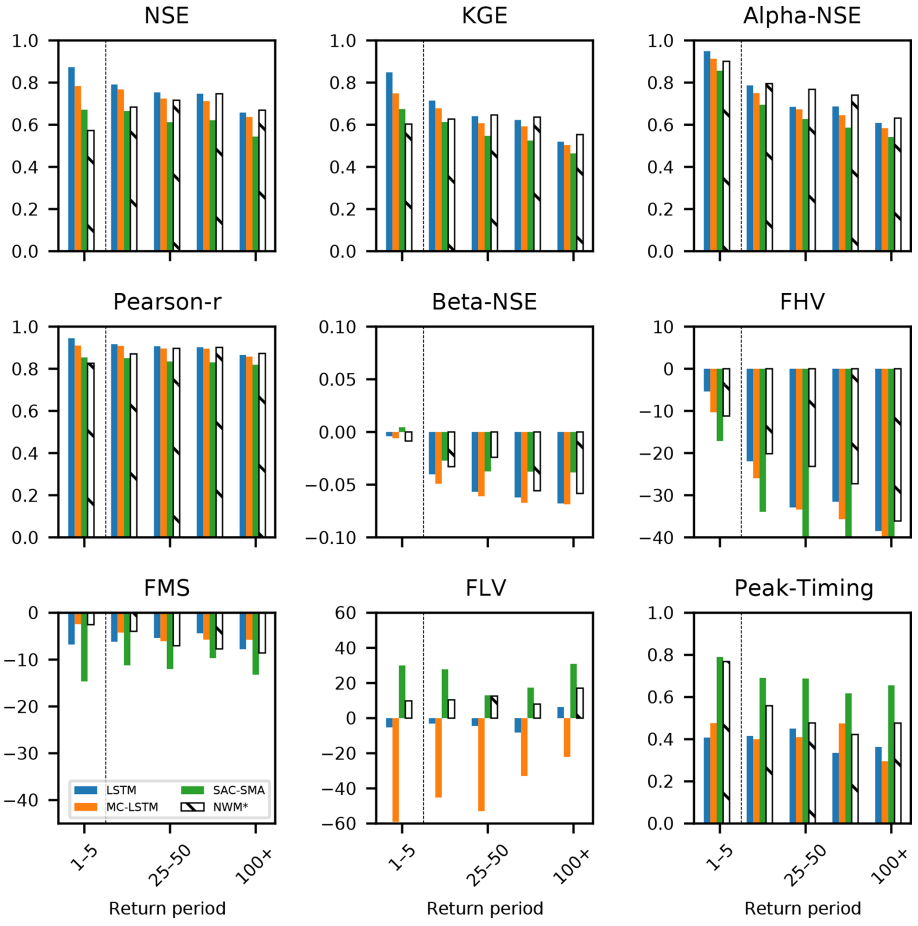

Figure C2. Metrics for the models trained only on high-probability years. The bins for return periods greater than 5 are out of sample for the LSTM, MS-LSTM and SAC-SMA models. The total number of samples in each bin is as follows: $n = 5969$ for 1–5, $n = 1260$ for 50–25, $n = 185$ for 25–50, $n = 91$ for 50–100 and $n = 84$ for 100+.