EGU23-1160, updated on 28 Mar 2024
https://doi.org/10.5194/egusphere-egu23-1160
EGU General Assembly 2023

Global streamflow modelling using process-informed machine learning

Michele Magni[1], Edwin H. Sutanudjaja[1], Youchen Shen[1,2], and Derek Karssenberg[1]
[1]Department of Physical Geography, Faculty of Geosciences, Utrecht University, Utrecht, the Netherlands
[2]Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands

Hydrological models include errors when reproducing real-world observations, due to uncertainties in their components that inevitably propagate to the simulated variable. A large body of research in streamflow prediction blends statistical learning into the hydrological sciences, modelling river discharge using meteorological variables and catchment attributes as predictors of observed streamflow.

We developed a novel hybrid framework that integrates information from the process-based global hydrological model PCR-GLOBWB to reduce prediction errors in streamflow simulations. Our statistical methodology employs simulated streamflow and state variables from PCR-GLOBWB as additional predictors of observed river discharge. These model outputs provide supplemental information that is effectively used in a random forest, trained on a global database of streamflow measurements, to improve estimates of simulated river discharge across the globe. PCR-GLOBWB was run for the years 1979-2019 at 30arcmin and daily resolution, and the simulated state variables were then aggregated to monthly time steps. A single random forest model was trained with these state variables, meteorological data and catchment attributes, as predictors of observed streamflow from 2286 stations worldwide.

Results based on cross-validation show that the model is capable of discerning between a variety of hydro-climatic conditions and river flow dynamics, improving KGE of PCR-GLOBWB simulations at more than 80% of testing locations and increasing median KGE from -0.02 in uncalibrated runs to 0.52 after post-processing. Performance boosts are usually independent of availability of streamflow data at a particular station, thus making our method a potential candidate in addressing prediction in poorly gauged and ungauged basins.

Further research is still needed to test the potential influence of additional predictors describing catchment and time-series behaviour. Cluster analysis is required to understand why the post-processing framework still performs poorly at some stations. For prediction purposes, future efforts should also be directed at testing the model at higher spatial resolutions globally, and at finer temporal resolutions.