CC comment (Sascha Ruzzante)

This is a useful and timely paper. The testing of model generalizability follows best practices in the hydrologic literature, but I'd like to take this opportunity to ask exactly what is meant by 'generalizability'. Is it:

a) Which model has the highest accuracy in an unseen warm test period?

b) Which model has the smallest reduction (or largest increase) in accuracy when moving from calibration (cold) to testing (warm) periods?

c) Which model most accurately simulates the change in hydrologic conditions between warm and cold periods?

Thank you for your comments. We define generalizability by conditions a + b. In our opinion, condition c alone is not a good proxy for generalizability, and can only be used as an extension of a and b.

These three alternate definitions are subtly different and suggest different statistical tests. This study relies on definitions (a) and (b). In climate change projection studies, however, it is common to summarize results as a percentage change from historical conditions, for which definition (c) is the most relevant.

Ideally, all three conditions apply for a model “suitable” for climate change assessments. c) alone doesn’t make a model generalizable. The model can get the change right for the wrong reasons. However, if conditions a+b are fulfilled, then it is likely that the change is also modelled correctly.

As an illustrative example: in a catchment, suppose the observed peak flow increases by 50%, from 100 cms to 150 cms, between the cold and warm periods. Models A and B give the following results:

Period	Observed	Model A	Model B
Cold	100	90	90
Warm	150	135	148
Change	50%	50%	64%

Model A has a persistent bias of -10%, and correctly predicts an increase in peak flows of 50%, while model B overpredicted the increase (64%). However, by definitions (a) and (b), we would select model B as the most generalizable since its accuracy in the warm period is highest and the accuracy improves from the calibration (cold) to the testing (warm) period.

We don’t fully agree. According to definitions a+b we would still choose model A, because generalizable models should 1) have good initial performance, i.e. models A and B would be comparable (i.e. both have a -10% bias), and 2) the model performance should not change too much between periods (i.e. model A would meet this criteria with a constant bias of -10%, no change in performance). Following these criteria, a model that shows a significant increase in performance will also fail the stability criteria (b) and the change in performance would indicate that the model's performance is tied to external factors, e.g. temperature.

This example is relevant to Table C1, where (for example) the hybrid model is shown to have the best performance for DVPB in the warm period (4.9%), but this represents a large reduction from the cold period DVPB (9.5%). In comparison, the LSTM has the most stable DVPB across the three periods, as indicated at L358. In this case, it seems that the LSTM will predict the *change* in DVPB best, and be most generalizable by definition (c). These numbers would, however, be more informative if compared on a catchment-by-catchment basis rather than comparing the median values.

In all figures, the catchment-by-catchment changes in performance are summarized in the boxplots. Summarizing the results in the form of boxplots was a pragmatic choice because we needed to find a way to quantify the overall behavior of the different models across the 918 catchments in our dataset. Alternatively, we could have shown scatterplots for specific pairs of models, which would have made it difficult to visualize differences between models.

I recommend including (maybe in an appendix) a comparison of the observed and simulated change in various hydrologic signatures between the cold and warm periods (eg., the mean annual flow, the mean monthly flow for each month, and the high flow, low flow, and drought metrics already calculated in the paper).

Thank you for these suggestions. We followed your recommendation and created a new figure quantifying the modelled changes in key streamflow metrics between the warm (warmest) and the cold period and comparing these changes to the observed ones. This allows us to assess not only whether model performance changes across periods (definitions a+b), but also whether each model appropriately captures the magnitude and direction of change in the respective metric in response to mainly changes in temperature (definition c). It seems that the conceptual and hybrid model can better capture the observed changes between the warm and cold periods. See figure below.

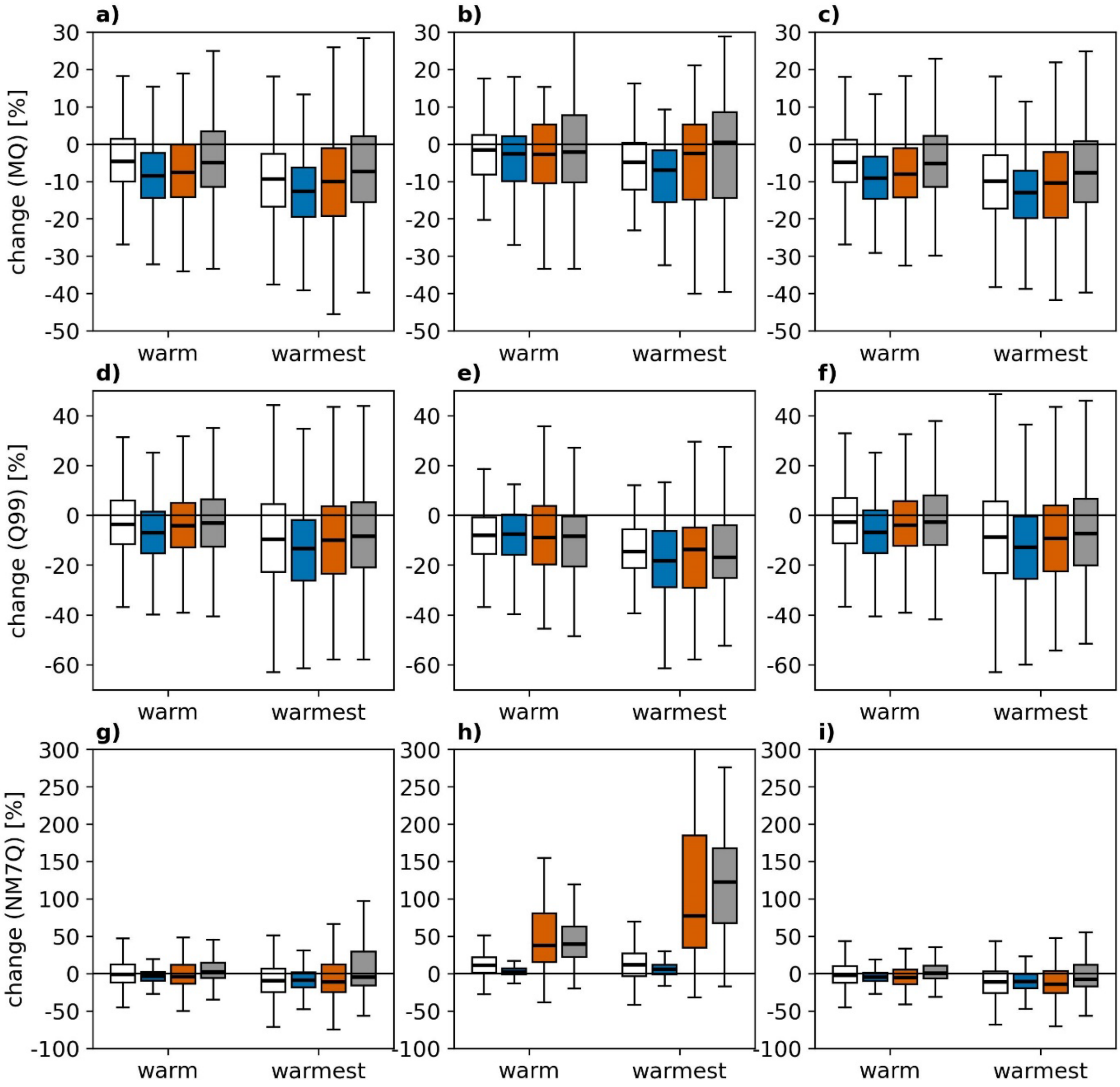

Figure x: Comparison of modelled and observed changes in streamflow metrics between the warm (warmest) and cold periods. Boxplots show catchment specific relative changes in (a-c) mean streamflow, (d-f) mean annual high flows (99th percentile), and (g-i) mean NM7Q between the warm (warmest) and cold period. The first column of panels (a, d, g) shows changes for all catchments, the center column (b, e, h) for snow-dominated catchments (more than 30% snowfall fraction and minimal glacier contribution), and the right column (c, f, i) for rainfall-dominated catchments (less than 30% snowfall fraction). Colors indicate: observations (white), LTSM (blue), Hybrid (orange), and HBV (grey).

As a second point, the LSTM is found to generalize most poorly in the warmest catchments. To me, this makes sense, given that the LSTM is extrapolating most strongly in these catchments. In the colder catchments (warm period), the LSTM can learn from the warm catchments (cold period). For the warm catchments (warm period) there is no analogue set of catchments from which to learn. It might be worthwhile to mention this explanation alongside the explanations already given (L427-442).

We agree that it is important to comment on this point, which we already do in *"l.462-66: Further, model generalizability could be improved by including catchments from regions whose climate is warmer than that of the region of interest as additional training catchments for deep learning based models (Martel et al., 2025). This allows these models to learn from a more diverse set of climates, which could help improve generalizability. Studies testing this approach on LSTMs have found that such geographically extended models can provide more realistic projections under climate sensitivity analyses than regional LSTMs (Wi and Steinschneider, 2022, 2024)."*