CH-RUN: A data-driven spatially contiguous runoff monitoring product for Switzerland; **https://doi.org/10.5194/egusphere-2024-993**

Response to anonymous reviewer #1; https://doi.org/10.5194/egusphere-2024-993-RC1

Comment 1

My first comment is about the paper title. I expected a somewhat different study based on the title. The mentioning of a runoff monitoring product in the title suggests some type of derived data product, though the focus of the manuscript is the extensive development of neural network approaches to perform the reconstruction. I assume that this is when the authors say data-driven, though I do think a title that better represents the actual study content would be preferable.

We agree that the title could be misleading and decided to change it to: "CH-RUN: A deep-learning-based spatially contiguous runoff reconstruction for Switzerland"

Comment 2 and 3

[2] The authors' view of "traditional hydrological models" is overly narrow (lines 24ff.). While physically-based (pb) models, like the one previously developed for Switzerland (PREVAH), have a high computational demand and are rather data hungry, this is not the case for all hydrological models. In fact, much of hydrology uses rather parsimonious models (GR4J, HyMod, PDM…) which do not put a high demand on computational resources. It would be good if the authors either refine their statement to pb models or widen it to include a wider range of model complexities. Given that the simulation of daily runoff is done with such simpler models in many countries, I would suggest the latter.

[3] A similar point can be made about the data need of hydrological models which is discussed in lines 36ff. Several widely used hydrological models can be driven by precipitation and temperature only – if they are of the parsimonious type

Overall, we agree with these comments. There are indeed less computationally demanding and data-hungry hydrological models. Such fast models are usually calibrated per catchment, and then regionalized (and so is PREVAH). While the specific data requirements and processing speed differs vastly among models, the end-to-end deep learning-based approach used here encompasses all of the mentioned advantages. Compared to PREVAH, which is used widely in scientific context to run scenarios and projections, our approach is fast and data-efficient.

We added a short note on different types of hydrological models to the introduction:

Traditional hydrological models offer pivotal insights into land-surface processes. For Switzerland, a diverse array of hydrological models has been employed (Horton et al., 2022), ranging from complex ones, which are heavily founded on physical principles, to lightweight ones using conceptual process representations with calibrated parameters. While the former offer detailed insights and control, they rely on a large number of inputs and are computationally expensive. The latter, in contrast, can be parsimonious in terms of data and computational resources, yet they need to be calibrated per catchment, which limits their applicability to prediction in ungauged catchments. The generalization to ungauged catchments via regionalization is possible, but introduces another layer of complexity (Beck et al., 2016). As a complementary approach, deep learning holds potential as a tool for hydrological modeling, both in terms of performance and efficiency (Nearing et al., 2021), and it comes with built-in regionalization when trained on multiple catchments jointly (Kratzert et al., 2024).

With this paragraph, we wanted to calrify that process-based models can be fast and data-efficient as well. We also made the following changes:

Instead of “reduced data requirements”, we now write “low data requirements”.

Furthermore, the low data requirements and computational efficiency of our model pave the way for simulating diverse scenarios and conducting comprehensive climate attribution studies. This represents a substantial progression in the field, allowing for the analysis of thousands of scenarios in a time frame significantly shorter than traditional methods.

We added “reduced data needs [...] *compared to the PREVAH model*.”

The reconstruction of runoff back to the early 1960s for Switzerland is a novelty enabled by the reduced data needs of our deep learning-based approach compared to the PREVAH model. Here, we evaluate the plausibility of the simulated patterns based on Fig. 7-9 by contrasting them to prior knowledge.

The overall trend towards drier conditions simulated by our data-driven model aligns with independent studies. This has

Comment 4

The authors use squared error metrics for model calibration. They then use disaggregated components of such metrics for further analysis- which I like. What I missed in the analysis is any assessment of whether these components show any structure across Switzerland. For example, Gudmundsson et al. (2012, WRR) showed for example a strong correlation between bias errors and elevation differences for some comparable catchments to those used here. Did you look for any systematic biases (in the context of Fig. 4)?

We agree that such an analysis would be interesting and decided to add a figure on spatial errors (new Figure 4, see below). With this figure, we can discuss spatial patterns of error components and link them to catchment properties. We also added the difference in the error components (LSTM minus PREVAH), to better understand how the models compare on a spatial basis. We also performed an explorative analysis to better understand the relationship between model performance (difference) and catchment properties.

The new figure:

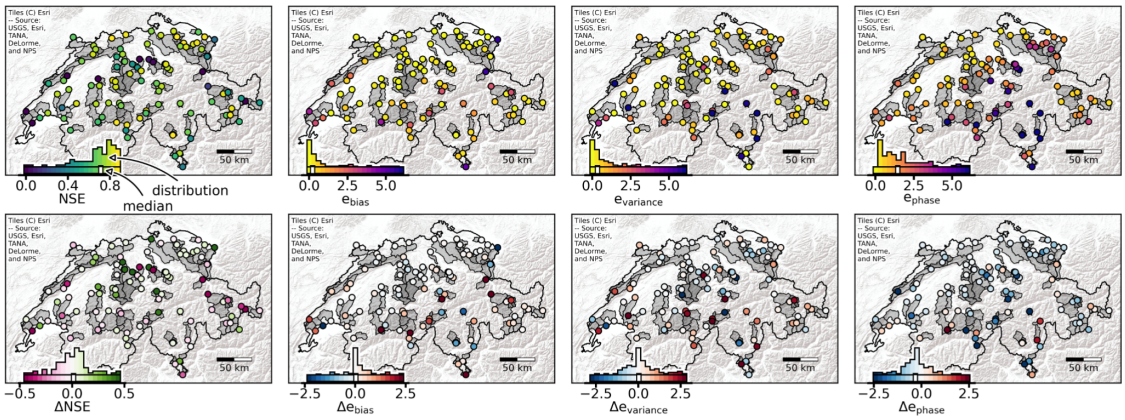

Figure 4. Spatial catchment-level performance of our best-performing model ("$\text{LSTM}_{\text{best}}$") contrasted to the PREVAH hydrological model. The top row shows the performance of $\text{LSTM}_{\text{best}}$, with the Nash-Sutcliffe modeling efficiency (NSE) in the most-left panel, and bias (e_{bias}), variance (e_{variance}), and phase (e_{phase}) error, in the remaining panels. Note that in the top panel, yellowish colors indicate better performance, i.e., for NSE, a larger number is better and for the error components, lower numbers are preferred. The bottom row shows the performance difference of $\text{LSTM}_{\text{best}}$ minus PREVAH. Here, reddish colors indicate that PREVAH performs better than $\text{LSTM}_{\text{best}}$, which is, for NSE, negative values, and for the error components, positive values. The inset histograms represent the distribution of catchment metrics, and the white bar indicates the median of the distribution, per panel. The evaluation is performed on the test set, but all catchments are in this set once in our cross-validation setup.

The updated results:

Next, we investigate the spatial distribution of the errors. First, we notice that the performance of $\text{LSTM}_{\text{best}}$ in terms of NSE, shown in the top-left panel Fig. 4, does not exhibit a clear spatial pattern. Yet, the model seems to struggle with some particular catchments. Interestingly, these are the very catchments where PREVAH outperformed $\text{LSTM}_{\text{best}}$ clearly (compare Fig. 4 top-left panel dark blue values to lower-left panel dark red values).

To understand how this spatial patterns are linked to catchment properties, we performed an exploratory analysis. First, we identified the tails of the distributions (inset histograms in Fig. 4, and PREVAH performance, not shown) using the 10th and 90th percentiles. We then compared properties of catchments in the tails to the "normal" group (between the 10th and 90th percentiles) using the two-sided, non-parametric Mann-Whitney U test with a significance level of $\alpha = 0.1$ (Mann and Whitney, 1947). The analysis was restricted to a subset of catchment properties: mean and variance of runoff, elevation, catchment area, and water body fraction. Here, we report the most notable findings from this ad-hoc analysis.

For $\text{LSTM}_{\text{best}}$, poor performance (NSE below 0.14) was observed in catchments with low runoff mean and variance, whereas good performance (NSE above 0.86) was achieved in catchments with high runoff variance. Similarly, PREVAH struggled (NSE below 0.08) in catchments with low runoff mean and variance, as well as in low-elevation and lake-dominated conditions, but performed well (NSE above 0.84) in catchments with high runoff mean, large catchment areas, and minimal lake presence. As expected, the bias of $\text{LSTM}_{\text{best}}$ was low in catchments with low runoff mean and variance, with variance error increasing in high runoff variance conditions. Phase error for $\text{LSTM}_{\text{best}}$ was lowest in catchments with low runoff mean and variance and large catchment area.

The updated discussion:

From the analysis of the spatial patterns of the model performance (Fig. 4), we learned that the $LSTM_{best}$ encountered challenges with dry catchments that have both low runoff mean and variance. This was not surprising due to the high signal-
410 to-noise ratio in runoff observations and the sensitivity to minor variability in the meteorological variables and catchment properties in dry catchments. Similarly, PREVAH struggled with dry conditions, but still it performed clearly better under such conditions. In contrast, $LSTM_{best}$ represents lake-dominated catchments with low elevation significantly better. This was expected, as PREVAH does not represent lake processes, and therefore, it cannot properly represent their dampening effect. The interaction with elevation could be explained by the fact that the largest lakes in Switzerland are at medium-to-low elevation.

Significant differences in NSE performance between the two models were observed in catchments with low runoff variance. PREVAH outperformed $LSTM_{best}$ (NSE improvement above 0.28) in catchments with both low runoff mean and variance. Conversely, $LSTM_{best}$ clearly outperformed PREVAH (NSE improvement above 0.28) in low-elevation, lake-dominated
290 catchments that also had low runoff variance.

Comment 5

The relative NSE range shown in the legend of Figure 3 seems very small. Is the variability shown in the various small plots actually relevant?

We will move this figure and its discussion to the appendix to make space for the more relevant figure from comment [4].

Comment 6

The other reviewers already made some comments and suggestions regarding the trend analysis performed, and the need to test a non-parametric strategy. I will not repeat his points in this review, but I believe that they are justified.

We agree with this criticism of the trend analysis. We followed the suggestion of the other reviewer and used Sen's slope to compute the robust trend per catchment. To reduce the impact of outliers for the comparison of observed and simulated trends, we use robust estimates for regression and Spearman's rank correlation.

With this robust analysis, the difference between CH-RUN and PREVAH are slightly reduced, but still, CH-RUN reproduces linear trends better with rank correlation ρ=0.6, compared to PREVAH with ρ=0.42. Both models appear to reproduce the trends less accurately compared to the previous analysis.

The updated results:

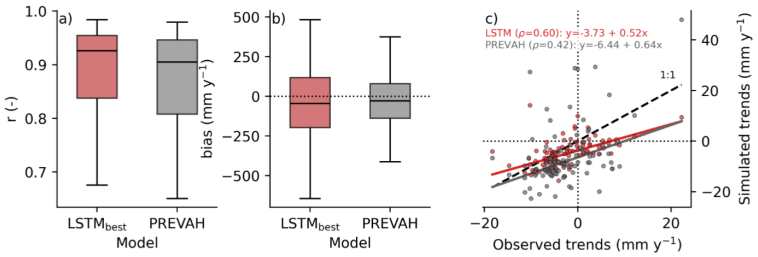

Figure 5. Catchment-level evaluation at the annual scale. a) The Pearson correlation (r) and b) bias in $\mathrm{mm\,y^{-1}}$ distribution across 98 training catchments evaluated on the test set. c) The simulated annual runoff trends compared to observations. The points represent the linear trend (found by robustified least squares fit) of single catchments. Note that for the trend calculation, the time range from 1995 (start of first test period) to 2020 (end of second test period) was used. The inset equation shows the linear least square fit and the corresponding rank correlations.

For a reconstruction product, it is crucial to adequately represent yearly variability and long-term trends. We, therefore, evaluate this aspect on annual runoff aggregates (Fig. 5). The best-performing model, $\mathrm{LSTM_{best}}$, represented the interannual variability (Fig. 5a), quantified as the Pearson correlation coefficient between the annual values for each catchment, well with a median of $r = 0.93$, and with 75 % of catchments above $r = 0.85$. The bias averages close to zero and for 50 % of the catchments, it was in the range of -250 to 250 $\mathrm{mm\,y^{-1}}$ (Fig. 5b). On the interannual variability, PREVAH showed a slightly lower correlation (Fig. 5a) across catchments with a median of $r = 0.91$. In terms of bias, PREVAH performed marginally better with a median closer to zero and a lower spread (Fig. 5b).

Figure 5c illustrates how the models captured spatial patterns of annual trends between January 1995 and December 2020 (Fig. 5c). The agreement was calculated by first computing the catchment-level linear trends for the observations and the simulations by PREVAH and $\mathrm{LSTM_{best}}$ independently using the robust Theil-Sen estimator (Sen, 1968). Then, we fit a regression between the observed and estimated trend slopes by the two models using robust regression with Huber weighting and the default tuning constant of $c = 1.345$ (Huber and Ronchetti, 2009). This approach reduces the impact of outliers by giving lower weight to large residuals. For quantifying the alignment of the simulated trends, we us Spearman correlation (ρ), which is relatively robust against outliers. While the $\mathrm{LSTM_{best}}$ represented the spatial patterns of the linear trend relatively well with a correlation of $\rho = 0.60$, PREVAH achieveed a correlation of $\rho = 0.42$. Both models underestimated the strength of negative and positive trends with slopes of 0.52 ($\mathrm{LSTM_{best}}$) and 0.64 (PREVAH), and they exhibited small negative biases of -3.73 ($\mathrm{LSTM_{best}}$) and -6.44 (PREVAH) $\mathrm{mm\,y^{-1}}$.

Comment 7

Rather than the qualitative evaluation in section 4.2, is there not enough information in the 98 catchment differences to show where and when PREVAH is better/worse?

See answer to comment [4].

Comment 8

Section 5.3 "The reconstruction of runoff back to the early 1960s for Switzerland is a novelty enabled by the reduced data needs of our deep learning-based approach." But would the NN benefit from additional data?

Yes, the neural networks would benefit from additional covariates. We did not systematically test this, but in preliminary model runs we found that adding more covariates (radiation, vp, tmin, tmax) helped. The differences were, however, not substantial. The dependency on air temperature and precipitation alone allowed us to extend the reconstruction back to the 1960s, which enabled the monitoring of long-term trends. Testing the capabilities of deep learning approaches in the time domain in more data-abundant periods has been done before and is out of the scope of this study.

Comment 9

(lines 446ff.) The authors state that "A limitation in our approach was the reliance solely on air temperature and precipitation data for long-term reconstruction, excluding other meteorological factors like cloud-related effects, which could only be indirectly approximated by the model." Can you name examples of hydrological models that consider cloud-related effects? Do you mean the consideration of sunshine hours? You could have used such information, couldn't you?

Yes, we meant cloud effects on radiation. The latter is used in many hydrological models (e.g., SWBM, PREVAH). The deep learning models can, in principle, learn such effects implicitly (precipitation means clouds means less radiation). We could use such data, yes, but sunshine hours are only available from the 1970s from MeteoSwiss. We changed the wording and hope that it is clearer now:

> A limitation in our approach was the reliance solely on air temperature and precipitation data for long-term reconstruction, excluding other meteorological factors like sunshine hours, which can only be implicitly approximated by the model via the available input variables. The assumption of static variables, such as land use and glacier coverage, being constant over time
> 470 is a necessary simplification but introduces potential inaccuracies. This is particularly critical as land use can vary and glacier areas are known to decrease over time, potentially leading to biases, especially in the early stages of the reconstruction where observational data are sparse.

Comment 10

Also, the authors state that the "The assumption of static variables, such as land use and glacier coverage, being constant over time is a necessary simplification but introduces potential inaccuracies." I am not completely clear why this is a necessary simplification. Why can changing forest cover and a limited contribution of melting glaciers not be included?

In principle, we could use such data, but it is difficult to find high-quality and harmonized historical data on land use or glaciers covering the entire period back to the 1960s. Even if such data is available in some form, the model architecture would need to be adapted to deal with inconsistent resolution etc. Thus, we consider this an interesting suggestion, but unfortunately out of the scope of this study.

Comment 11

Conclusions: "One of the major strengths of our approach lies in its computational efficiency, which opens up possibilities for contiguous near real-time monitoring and potentially forecasting of runoff." And "…allowing for the rapid evaluation of thousands of scenarios that were not feasible with traditional physically-based models." Here the authors state their assumption of "traditional physically-based models" which is not the same as traditional hydrological models. It would be good to clarify this difference in the Introduction section.

This should read "traditional hydrological models". We will clarify this in the revised manuscript. The broader discussion about model types and their strengths and weaknesses is covered in the answers to comment 3 and 4.