Can discharge be used to inversely correct precipitation? (hess-2024-375)

Ashish Manoj J, Ralf Loritz, Hoshin Gupta, Erwin Zehe

The authors would like to thank Anonymous Reviewer 1 for carefully reviewing our manuscript and for providing their valuable comments and suggestions, which we believe will be very helpful in improving the overall structure and quality of the manuscript. The following responses have been prepared to address all the reviewers' comments point-by-point. We have responded (in black) to the reviewer's comment (in blue).

General comments:

The authors present a method to improve the estimation of catchment-average effective precipitation from the ERA5 product by utilizing the information contained in stream flow data and a regional LSTM model.

To validate this interesting approach the authors model the runoff using this catchment-average effective precipitation as forcing and compare it to the runoff in the CAMELS data set. Averaged over all catchments contained in the CAMELS data set this approach improves the modelled runoff compared to using only ERA5 precipitation estimates as forcing.

The paper is well written and has a reasonable length. However, I think the authors could address the topic of "scale" more in depth. This starts at describing the used data sets in more detail, especially by mentioning their spatial and temporal resolution. Furthermore, the selection of the out-of-sample data sets as "proof-of-concept" is restricted to very small-scale basins. It is especially at this scale that daily ERA5 precipitation will most likely not perform well as forcing for a hydrological model due to its coarse spatial resolution.

The authors state that the LSTM model is estimating catchment-average precipitation amounts. This implies that the introduced approach might not perform equally well for differently-sized basins, when the runoff dynamics shift from surface-runoff to baseflow dominated basins. In my opinion the authors should elaborate more on this topic.

The study is interesting and introduces a promising approach to improve gridded precipitation products which is why I recommend its publication in NHESS after addressing the following comments.

We thank the reviewer for constructive, supportive suggestions and for highlighting the work's potential. We have prepared the following points to address the main comments raised by the reviewer:

A. Although the out-of-sample catchments we selected are relatively small, our approach using the LSTM model—trained on much larger catchments—showed skill at adjusting the (under) estimated precipitation values for these events at such smaller scales. This is noteworthy given that ERA5 Land precipitation typically performs poorly at this scale; only about 9% of the catchments in our training dataset had areas smaller than 100 km^2. This finding demonstrates the ability of our inversion methodology and spatial transfer learning (He et al., 2011) to effectively leverage knowledge from data-rich, well-represented regions to make predictions in data-scarce areas.

B. To address the question of performance in differently sized basins, we will supplement the existing analysis by an evaluation of the methodology over larger, previously unseen test catchments from the recently published Caravan CAMELS-DE (Loritz et al., 2024) dataset. The results (Figure 1: camelsde_DEA11130) again indicate a reduction in relative errors in peak discharge and flood volume for such large catchments (>3000 km^2), using the inversely generated precipitation.

C. We agree with the reviewer that our usage of the term "effective precipitation" requires more clarity. To begin with, it is important to stress that precipitation uncertainty is rarely considered when quantifying model output uncertainty; while studies are usually conducted to show how differences in simulated discharge can be as a consequence of changing precipitation input, they rarely look at how much improvement of the model performance would be possible by using different but plausible precipitation (Bárdossy et al., 2022, 2020). "True" precipitation estimates are not known at the catchment scale. We obtain estimates of them (with considerable uncertainty) by either interpolating station data or averaging gridded data from reanalysis/remote sensing products. Our aim was to generate a precipitation time series (estimate) that is more "consistent" with the dynamics captured in the discharge record. We then benchmarked our new estimate by using forward hydrological models over the Elsenz Schwarzbach (Figures 6 & 7 in the original draft) and Lippe catchment (Figure 1 in this document). As our results indicate, the inversely generated precipitation estimate reduces both volume and peak errors for the HBV model simulation in both catchments.

We will update the manuscript to reflect these points and include the new analysis.

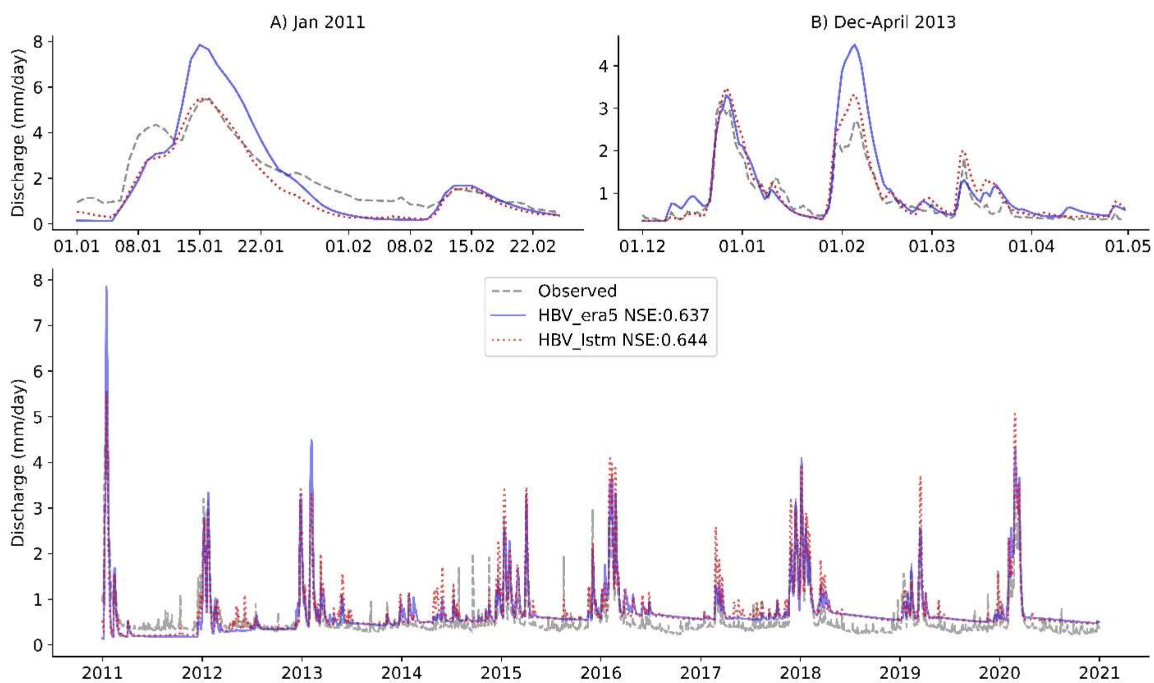

Figure 1 Observed and simulated runoff (using the HBV model) at the Lippe catchment (camelsde_DEA11130). The blue line denotes the streamflow simulated using the ERA5 Land precipitation product, while the red curve depicts the simulations using the inversely-estimate precipitation obtained using the regional LSTM model. Moreover, two rainfall-runoff events are highlighted and displayed.

Specific comments:

- The question about the feasibility of the introduced approach (as stated in the abstract) could be answered more clearly. While the authors show that their approach indeed improves the modelled discharge, they could elaborate more on how this method should be actually applied. E.g. a general improvement of products like ERA5: Does the method have potential to adjust the precipitation estimates of ERA5, e.g. as a pre-processing step? What could be general use-cases of the method?

We agree with the reviewer that more information about the proposed method's application could enhance the work's potential. Some of the main points we will add are:

A. Transfer learning to data-scarce regions: For a number of smaller catchments, no precipitation gauges exist within the catchment boundaries. Due to this, highly uncertain and erroneous precipitation estimates are often employed for hydrological modelling over these catchments, leading to the underrepresentation of high impact events such as convective storms. The inverse methodology could be applied to generate more realistic representations of extreme precipitation statistics at these scales, which could then be utilised to design flood defence measures.
B. Improvement of gridded products: Reanalysis data, by definition, are a mix of observations and past short-range weather forecasts rerun with modern weather forecasting models (ECMWF, 2023). The inversion technique could be used as another final layer of post-processing for the model outputs to ensure that the final product is more consistent with the variabilities observed in the discharge record. In line with this, it would also be interesting to use machine learning and other data-driven approaches to generate estimates of the spatial fields of precipitation (Bárdossy et al., 2022, 2020) conditioned on the discharge information.
C. Reconstruction of past floods: As a more general use case, the methodology could be applied to reconstruct information about the driving storms that caused some of the devastating past floods. There exists a wealth of hydrological information about such events (Bronstert et al., 2018; Seidel et al., 2009), either in the form of storm water level markings or observational flood records.

- The authors state that the method could be useful to "reproduce small-scale high-impact events" (l.15-16) which also reflects in their choice of out-of-sample catchments. However, in my opinion it should be stated more clearly (e.g. in the "Limitations") that using daily precipitation sums of products like ERA5 and EOBS should really be the last resort when trying to reproduce small-scale hydrological events (with or without LSTM correction). The selected flood events in four out-of-sample catchments were dominated by sub-daily rainfall and a rapid, probably also sub-daily, flood response that can just not be captured with daily sums and the spatial and temporal resolution of ERA5 and EOBS.

The reviewer has raised a well-justified concern about using coarse-resolution precipitation products to reproduce small-scale hydrological events. While it is not ideal to use only such products, the scarcity of real-world data and the rarity of these events sometimes necessitate a modelling decision to incorporate these coarser estimates. We will address this as described in our response to the main comment.

- In Figure 3 you are comparing ERA5 and EOBS precipitation estimates to the effective precipitation generated by the LSTM. Therefore, the lower amounts from the LSTM seem obvious. How useful is this comparison between two different types of precipitation?

True precipitation rates at the catchment scale are seldom accurately known. The various products used carry significant uncertainty, particularly when applied in rainfall-runoff models. The main goal of our analysis at the continental spatial scale was to highlight systemic biases and identify areas where further investigation could help address these discrepancies. We found that while the LSTM model consistently underestimated extreme values, the spatial gradients were still well represented.

- **Line 318**: *"this consistent underestimation also seems physically plausible"*. Do I understand correctly that the LSTM-model systematically underestimates the precipitation in order to generate less surface runoff, to compensate for lacking baseflow dynamics? The systematic underestimation of precipitation from the LSTM model (as seen in Fig. 3) seems problematic to me because then the inversely derived precipitation would be <u>only useful to model the discharge in</u>

surface-runoff dominated catchments. The four out-of-sample catchments are all quite small: How would the LSTM-derived precipitation perform when forcing a hydrological model in a larger out-of-sample catchment that is dominated by base-flow dynamics? Please clarify this, because in my view this is a key point. You could also consider splitting the results of the two regional LSTM models into groups (by basin size) to investigate whether the underestimation of precipitation is more problematic for larger catchments when modelling the discharge.

- If you use "catchment-average effective precipitation" as forcing for HBV and CatFlow, what happens with the soil water dynamics, and interception? Do they get subtracted again from the effective precipitation input? Maybe the overall wording is misleading and the LSTM is not really returning effective precipitation?

We thank the reviewer for demanding more clarity about the consistent underestimation by the LSTM model. We feel that this can be attributed to the following reasons:

a) The LSTM model looks for recurrence in patterns and mean conditions. This means that it can indeed account for consistent baseflow dynamics (as also indicated by our new analysis over the larger Lippe catchment, Figure 1). In extreme floods (Merz et al., 2021), the relative contributions of each component can vary significantly, depending on various factors such as the antecedent conditions of the catchment area. The model likely struggles to learn this variability while attempting to invert and obtain the driving precipitation values.
b) Given the non-linear nature of the inverse problem, there are always multiple possible solutions. Since the model is trained to minimize the mean squared error (Gupta et al., 2009), it may also tend to consistently predict lower values (on peaks) to effectively reduce the average error during training.
c) Recent studies have shown that the LSTM models have a theoretical upper limit for prediction based on weights and biases of the head linear layer (Espinoza et al., 2024; Kratzert et al., 2024). In simpler terms, irrespective of the input series, the predicted values can never exceed a theoretical limit

(which is established during the training phase). This so called 'saturation problem' (Chen and Chang, 1996; Rakitianskaia and Engelbrecht, 2015) of the LSTM architecture would also lead to the underestimation of some of the peak storm events.

We will discuss all these relevant points and opt for clearer wording concerning the precipitation estimate from the LSTM model in the revised version of the manuscript.

- The authors mention that their approach could be useful in "data-scarce regions world wide". However, the uneven distribution or lack of stream gauges could also introduce a bias to the LSTM model.

This is a valid point as data scarcity is a pertinent challenge in hydrological modelling, however, as shown in the IAHS PUB (Prediction in Ungauged Basins: Sivapalan et al., 2003) decade, there is much knowledge to be gained from spatial transfer learning. Concerning the LSTM model, we believe that while it will consistently perform better in regions similar to those it has already encountered during training, it also possesses some generalization capabilities. If these capabilities are implemented in a hydrologically sound way, they could significantly help address the ungauged basin problem.

- A very brief explanation of the used data sets in the data section would be helpful for the reader, even though most of these data sets are commonly used. The temporal and spatial resolution of these data sets has to be mentioned, as well as the way these data sets are derived (e.g. based on station data). The authors do this partially, e.g., in line 128, for the E-OBS dataset but I would recommend doing it in a more structured way, for instance in a table.

- CARAVAN
- ERA5
- E-OBS
- CAMELS

You could also consider mentioning MERRA and GLDAS already here. Additionally, you could consider splitting the "Data and Methods" section into "Data" and "Methods" to provide more clarity.

We will follow the reviewer's suggestion to have a dedicated Data section with a description of the different products. A new table detailing the spatial and temporal resolution of all the data sources used in the study will also be added.

- Please describe your measures of goodness briefly and why chose to use them: mean wet days, spearman lag, 95th percentile. For mean wet days you should also mention the unit mm, otherwise this gets confusing. This also applies to Figure 3 where the unit "mm" is missing

The goodness of fit measures will be explained in the Methodology section. The figure 3 will also be updated to show the units.

- **Line 121**: Please write out the abbreviations for the five catchment static attributes.

This will be added to the revised manuscript.

- **Line 96**: "the model was provided with a 7-day lead time series for discharge"

  Don't you also have to provide the 7-day lead time series of the other forcings? Can you explain why you decided for a lead time of 7 days? Is the lead time not catchment specific/scale dependent, or does 7 days cover everything?

The 7-day lead time was chosen as this provided a reasonable upper estimate for the time of concentration in the catchments. Since the objective was not on forecasting, we opted not to give the lead time values for the other meteorological variables. While the lead time values are indeed catchment-specific, since the catchment area is already given as a static attribute to the LSTM model, we expect the model to learn this dependency from the data.

Technical corrections:

Generally, the plots are not displayed nicely in the PDF. I am not sure if the reason is the processing by HESS or if you should increase the resolution.

We can confirm that we have thoroughly rechecked the quality of each individual figure. The decline in PDF quality is likely due to the conversion of embedded figures from the original Word document. This issue will be resolved during the final typesetting phase, as we will provide the original high-resolution images to HESS. We agree with all the technical corrections proposed by the reviewer and will incorporate the necessary changes into the revised draft.

Line 40: The reference "*Clerc-schwarzenbach et al.,2024*" should be written with a capital "S": Clerc-Schwarzenbach

Thank you for pointing this out. We will correct the reference in the revised version.

Line 45: The reference Manoj J et al. is missing a "." after "J".

This will be added.

Line 46: "*precipitation forcings data*". This sentence seems a bit awkward. How about. "precipitation forcing data"?

We will change this to *precipitation forcing data.*

Line 50: "*it is usually the rainstorm events occurring in poorly observed areas that lead to high impacts*": This sentence could be misunderstood in a way that high-impact-rainstorms preferably happen in areas <u>because</u> they are poorly observed, while in fact the observation network is too sparse and the majority of high-impact-rainstorms is simply not observed.

We appreciate the reviewer's point about the need for more clarity. This sentence will be rephrased to avoid any ambiguity.

Line 56-59: "*While the classical "forward rainfall-runoff generation problem" has received considerable attention over various decades (Montanari et al., 2013; Sivapalan et al., 2003), a smaller subset of studies (Brocca et al., 2013; Kirchner, 2009; Kretzschmar et al., 2014;*

Krier et al., 2012; Teuling et al., 2010) has investigated the feasibility of tackling the inverse problem more efficiently."

What do you mean with "more efficiently"? More efficient than what?

Thank you for pointing out this typo. The term 'more' will be omitted in the new manuscript.

Line 59: Reference in the wrong format: Kirchner (Kirchner, 2009) should be Kirchner (2009).

We will correct this in the revised draft.

Line 81: You could already make a reference to Table 2, to reveal more information about the out-of-sample catchments.

This will be added to the revised version.

Line 90: References in wrong format. Parenthesis should just be around the year. E.g. Kratzert et al. (2018).

Line 93: Reference in wrong format. Should be "Loritz et al., 2024".

The references will be corrected.

Line 104 and 105: Why are you using two different references for the HBV model?

The first reference is the more general publication introducing the HBV model, while the second refers to the specific version of HBV used in our study. We will standardise the references throughout the revised draft.

Line 156: You already spelled out the abbreviation of the HBV model previously and do not need to do it again here.

We will omit the repetition.

Line 184: "underlying causes of precipitation". What do you mean by this? Maybe "causes" is the wrong word?

This will be changed to 'driving precipitation' in the revised draft.

Line 190: I think you confuse "without_discharge" and "with_discharge" here. It looks like "with_discharge" has a steeper curve.

We agree with the reviewer on the ambiguity of the statement. This will be removed in the revised draft.

Figure 3: Please add the units, otherwise it is unclear whether it is [days] or [mm] for "mean (wet days)". You mention "mean wet days" before but explain the unit first in line 204.

We thank the reviewer for pointing out the inconsistency in our usage. The same will be corrected, and units will be added in the revised version.

Line 241: "*precipitation values revealed closer estimates to those reported*" -> I think it should be: precipitation values revealed estimates closer to those reported.

We will update this to *precipitation values revealed estimates closer to those reported.*

Table 2: The first column of the table is not correctly formatted.

The table will be reformatted in the revised draft.

Line 326-334: This paragraph seems unnecessarily long. If your main message is that products like ERA5 have a too coarse resolution to capture small-scale precipitation events you can shorten this. Instead, you could elaborate here on the potential of your method, improving such products.

We will completely rewrite this paragraph, adding information about the application of the proposed method and its potential to improve gridded products.

Line 374: "*adds its own biases to the modelling exercise*" Can you explain this?

Our goal was to emphasize that in data-driven models, the choice of training function and evaluation metric significantly influences the results, especially when we focus solely on the numerical values of the goodness of fit measure. In this study, we aimed to address this issue by relying less on the evaluation measure (NSE) and placing greater emphasis on the feasibility of our predictions through runoff coefficient analysis at the event scale.

We plan to discuss more on this in the revised version and will also include relevant references.

Figure S2 and S3: Minor detail: Maybe it would look better, if you put the north-arrow in some corner, so it is less prominent? In Figure 3 there is no legend, so you could at least mention the red triangles in the caption.

We will update both figures (to also show the new out-of-sample catchment -Lippe) and move the north arrow. The captions of Figure S3 will also be updated.

Figure S5: How did you derive the gridded precipitation? By interpolating the rain gauges? It is not recommended using the rainbow color map because it is not perceptually uniform and therefore not easy to distinguish for people with some kind of color blindness. See also the HESS guidelines, section "color schemes":
https://www.hydrology-and-earth-system-sciences.net/submission.html#figurestables

The gridded precipitation depicted in the figure is from a radar product (Kachelmannwetter, 2023) operating over the Elsenz Schwarbach. The original figure was taken from an earlier publication (Manoj J et al., 2024). We will update this figure to also include the new test catchment area over the Lippe.

References

Bárdossy, A., Anwar, F., Seidel, J., 2020. Hydrological Modelling in Data Sparse Environment: Inverse Modelling of a Historical Flood Event. Water (Switzerland) 12. https://doi.org/10.3390/w12113242

Bárdossy, A., Kilsby, C., Birkinshaw, S., Wang, N., Anwar, F., 2022. Is Precipitation Responsible for the Most Hydrological Model Uncertainty? Front. Water 4, 1–17. https://doi.org/10.3389/frwa.2022.836554

Bronstert, A., Agarwal, A., Boessenkool, B., Crisologo, I., Fischer, M., Heistermann, M., Köhn-Reich, L., López-Tarazón, J.A., Moran, T., Ozturk, U., Reinhardt-Imjela, C., Wendi, D., 2018. Forensic hydro-meteorological analysis of an extreme flash flood: The 2016-05-29 event in Braunsbach, SW Germany. Sci. Total Environ. 630, 977–991. https://doi.org/10.1016/j.scitotenv.2018.02.241

Chen, C.T., Chang, W. Der, 1996. A feedforward neural network with function shape autotuning. Neural Networks 9, 627–641. https://doi.org/10.1016/0893-6080(96)00006-8

ECMWF, 2023. Factsheet on Reanalysis [WWW Document]. 2023. URL https://www.ecmwf.int/en/about/media-centre/focus/2023/fact-sheet-reanalysis

Espinoza, E.A., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Chaves, Á., Bäuerle, N., Ehret, U., 2024. Analyzing the generalization capabilities of hybrid hydrological models for extrapolation to extreme events 1–17.

Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

He, Y., Bárdossy, A., Zehe, E., 2011. A review of regionalisation for continuous streamflow simulation. Hydrol. Earth Syst. Sci. 15, 3539–3553. https://doi.org/10.5194/hess-15-3539-2011

Kachelmannwetter, 2023. Kachelmannwetter [WWW Document]. URL https://kachelmannwetter.com/de

Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. Hydrol. Earth Syst. Sci 28, 4187–4201.

Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., Tarasova, L., 2024. CAMELS-DE: hydro-meteorological time series and attributes for 1555 catchments in Germany. Earth Syst. Dyn. Discuss. 1–30.

Manoj J, A., Loritz, R., Villinger, F., Mälicke, M., Koopaeidar, M., Göppert, H., Zehe, E., 2024. Toward Flash Flood Modeling Using Gradient Resolving Representative Hillslopes. Water Resour. Res. 60. https://doi.org/10.1029/2023WR036420

Merz, B., Blöschl, G., Vorogushyn, S., Dottori, F., Aerts, J.C.J.H., Bates, P., Bertola, M., Kemter, M., Kreibich, H., Lall, U., Macdonald, E., 2021. Causes, impacts and patterns of disastrous river floods. Nat. Rev. Earth Environ. 0123456789. https://doi.org/10.1038/s43017-021-00195-3

Rakitianskaia, A., Engelbrecht, A., 2015. Measuring saturation in neural networks. Proc. - 2015 IEEE Symp. Ser. Comput. Intell. SSCI 2015 1423–1430. https://doi.org/10.1109/SSCI.2015.202

Seidel, J., Imbery, F., Dostal, P., Sudhaus, D., Ëurger, K., 2009. Potential of historical meteorological and hydrological data for the reconstruction of historical flood events-the example of the 1882 flood in southwest Germany. Nat. Hazards Earth Syst. Sci. 9, 175–183. https://doi.org/10.5194/nhess-9-175-2009

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J.J., Mendiondo, E.M., O'Connell, P.E., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. Hydrol. Sci. J. 48, 857–880. https://doi.org/10.1623/hysj.48.6.857.51421